

Hideyuki Azegami

Shape Optimization Problems

Springer Optimization and Its Applications

Volume 164

Series Editors

Panos M. Pardalos , *University of Florida*

My T. Thai , *University of Florida*

Honorary Editor

Ding-Zhu Du, *University of Texas at Dallas*

Advisory Editors

Roman V. Belavkin, *Middlesex University*

John R. Birge, *University of Chicago*

Sergiy Butenko, *Texas A&M University*

Franco Giannessi, *University of Pisa*

Vipin Kumar, *University of Minnesota*

Anna Nagurney, *University of Massachusetts Amherst*

Jun Pei, *Hefei University of Technology*

Oleg Prokopyev, *University of Pittsburgh*

Steffen Rebennack, *Karlsruhe Institute of Technology*

Mauricio Resende, *Amazon*

Tamás Terlaky, *Lehigh University*

Van Vu, *Yale University*

Guoliang Xue, *Arizona State University*

Yinyu Ye, *Stanford University*

Aims and Scope

Optimization has continued to expand in all directions at an astonishing rate. New algorithmic and theoretical techniques are continually developing and the diffusion into other disciplines is proceeding at a rapid pace, with a spot light on machine learning, artificial intelligence, and quantum computing. Our knowledge of all aspects of the field has grown even more profound. At the same time, one of the most striking trends in optimization is the constantly increasing emphasis on the interdisciplinary nature of the field. Optimization has been a basic tool in areas not limited to applied mathematics, engineering, medicine, economics, computer science, operations research, and other sciences.

The series **Springer Optimization and Its Applications (SOIA)** aims to publish state-of-the-art expository works (monographs, contributed volumes, textbooks, handbooks) that focus on theory, methods, and applications of optimization. Topics covered include, but are not limited to, nonlinear optimization, combinatorial optimization, continuous optimization, stochastic optimization, Bayesian optimization, optimal control, discrete optimization, multi-objective optimization, and more. New to the series portfolio include Works at the intersection of optimization and machine learning, artificial intelligence, and quantum computing.

Volumes from this series are indexed by Web of Science, zbMATH, Mathematical Reviews, and SCOPUS.

More information about this series at <http://www.springer.com/series/7393>

Hideyuki Azegami

Shape Optimization Problems



Springer

Hideyuki Azegami
Graduate School of Informatics
Nagoya University
Nagoya, Aichi, Japan

ISSN 1931-6828 ISSN 1931-6836 (electronic)
Springer Optimization and Its Applications
ISBN 978-981-15-7617-1 ISBN 978-981-15-7618-8 (eBook)
<https://doi.org/10.1007/978-981-15-7618-8>

© Springer Nature Singapore Pte Ltd. 2020

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721,
Singapore

Preface

Three-dimensional modeling and numerical analysis utilizing computers are conducted on a day-to-day basis in product design, and there is a growing interest in the optimization of design using the results of numerical analysis. Optimization problems in which geometrical parameters of a three-dimensional model defined on a computer are taken to be the design variables are called parametric shape optimization problems. Software for solving such problems is constructed based on experimental design methods or mathematical programming. However, when increasing the number of design variables in order to increase the degrees of freedom, the solution rapidly becomes more difficult to find.

Conversely, problems in which no geometrical parameters are used and optimum shapes are obtained from an arbitrary form are called nonparametric shape optimization problems. In particular, problems in which the optimum shape is sought by introducing holes are called topology optimization problems. Moreover, a problem in which the optimum shape is obtained through domain variations is referred to as a shape optimization problem of domain variation type or a shape optimization problem in a restrictive sense. Numerical methods have been developed to solve these problems and are being used to seek practical optimum shapes.

Figure 1 shows a numerical result for a topology optimization problem in maximizing the rigidity of a heel counter. Chapter 8 explains in detail how to choose the design variables. The external force estimated by the experiment is assumed to be known, and the work done by it, which is defined as mean compliance in the body of this book, is chosen as the objective function. The condition that mass does not exceed a prescribed value is posed as a constraint condition. Figure 2 shows the result of a numerical analysis with respect to a shape optimization problem of domain variation type aiming to decrease the weight of an aluminum wheel. Chapter 9 explains how the design variables should be selected in this case, where volume is selected as the objective function. A constraint is imposed on the Kreisselmeier–Steinbauer function, which expresses the maximum value of Mises stress in an integral form so that it does not exceed its initial value. The analysis also includes a constraint on the shape variations which take into account the symmetry and manufacturing requirements of the model.

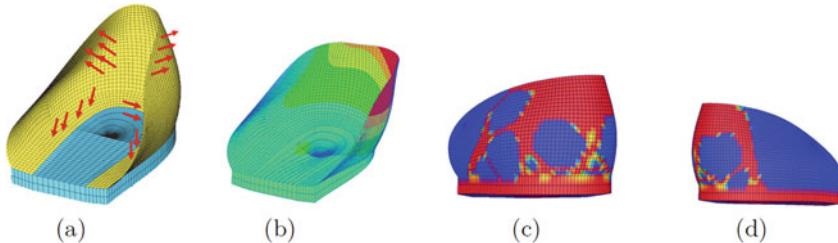


Fig. 1 Rigidity maximization of a heel counter (provided by ASICS Corporation). (a) External force and fixed sole. (b) Mises stress initial model. (c) Optimized density (inside). (d) Optimized density (outside)

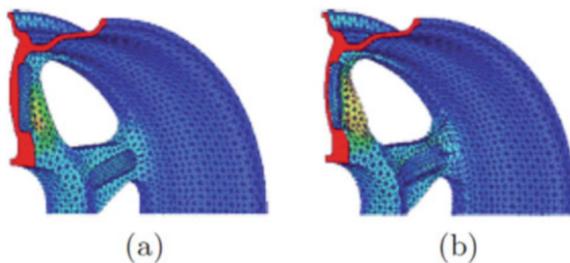


Fig. 2 Lightening of an aluminum wheel (provided by Quint Corporation). (a) Initial shape. (b) Optimized shape

The fundamental principles of nonparametric shape optimization programs used to obtain these results are also based on mathematical programming. However, from the fact that the design variables are functions expressing densities or domain variations, there are issues that cannot be dealt with by finite-dimensional vector spaces, which are the platform for mathematical programming. In this regard, it is possible to drop the nonparametric shape optimization programs into parametric optimization problems in finite-dimensional vector spaces by discretizing continua using a procedure such as the finite element method. However, when using these methods, one faces a new problem that there is an insufficient smoothness of the function used in updating the density or domain variation. This problem emerges as numerical instability phenomena when conducting numerical analysis. In order to solve this problem, there is a need to think of solutions based on theories capturing shape optimization problems as function optimization problems. The numerical techniques described above are based on such theories. However, there are no books explaining such theories from their foundations.

This book explains the formulation and solution of shape optimization problems for continua such as elastic bodies and flow fields in detail from the basics, bearing in mind readers with an engineering background. A continuum refers to a domain over which a boundary value problem of a partial differential equation is defined. With respect to the partial differential equation, considering a static elastic body or

a steady flow field, elliptic partial differential equations are assumed. The theories shown in this book, however, can be applied to time-dependent initial boundary value problems related to hyperbolic or parabolic partial differential equations, as well as to nonlinear problems. These results will be introduced on another occasion. Hence, the shape optimization problems dealt with in this book are described as follows. We first describe a boundary value problem of an elliptic partial differential equation as a state determination problem and then define the state determination problem so that when a function of the design variable expressing the density or domain variation is given, it has a unique solution. Using the design variable and its solution, we define several cost functions by boundary integrals or domain integrals. Among these cost functions, one is set as the objective function and the remaining as constraint functions in order to construct an optimum design problem. In this way, throughout this book, shape optimization problems will be constructed in the framework of function optimization problems. Their solutions will also be considered as solutions to the function optimization problems.

Based on this sort of conception, this book uses a structure as described below. Chapter 1 examines a simple optimum design problem of a one-dimensional continuum and looks at the process until optimum conditions are obtained. In this chapter, the Lagrange multiplier method (adjoint variable method) is used without proof. After understanding the usage of the Lagrange multiplier method, optimization theories will be studied from the basics in Chap. 2. It is desired to have an understanding of the principle of the Lagrange multiplier method in this chapter. After mastering optimization theories, in Chap. 3, algorithms for obtaining the optimal solutions will be considered. The theories and algorithms shown here are the fundamentals of the academic field referred to as mathematical programming. However, these are applicable to optimization problems defined on finite-dimensional vector spaces. In order to tackle optimization problems defined on functional spaces constructed from sets of functions, which is the aim of this book, functions need to be treated as vectors. A theory systematizing its use is functional analysis. In Chap. 4, the variational principle in mechanics is used to exemplify the basic thinking and results regarding the functional analysis.

After such preparations, Chap. 5 will look at boundary value problems of partial differential equations which are the platform of this book. Chapter 6 will look at methods for performing numerical analyses of such problems. Numerical analysis is one of the academic fields which continues to develop, even now, and has a variety of perspectives for study. In this book, because theories such as error estimation are established, finite element methods using the Galerkin method as a guiding principle will be looked at. Here, remarkable results from the functional analysis will be used to show the unique existence of solutions and in error estimations.

With these results in mind, in Chap. 7, an optimum design problem with a level of abstraction, which can be used in the unification of shape optimization problems considered in this book, is defined, and its solution and algorithms are considered. The solutions and algorithms have the same framework as those shown in Chap. 3. Here, however, vector spaces with respect to design variables are replaced by function spaces. In Chaps. 8 and 9, the abstract optimum design problem

is translated into topology optimization problems of density variation type and shape optimization problems of domain variation type, respectively. In both chapters, the details of the theory are shown using the Poisson problem, for simplicity. We finally consider shape optimization problems for a linear elastic body and a Stokes flow field, which are important in engineering, where a method to obtain the derivatives of cost functions will be looked at in detail.

Based on the above synopsis, this book will build up theorems using results in mathematics. Therefore, we shall be summarizing the key points in mathematical definitions and theorems. This method of expression is advanced by the fact that the provable facts are clearly shown. If something that we want to investigate in the future is contained in the framework of mathematics, setting up a theory using theorems prepared by great mathematicians is thought to be an extremely effective method. Conversely, mathematics attempts to heighten the level of abstractness in order to understand many things in a unified fashion. This characteristic is also the reason that it can baffle readers with an engineering background. Hence, an attempt has been made to provide explanations using examples from dynamics with the aim of accurately denoting the provable facts using definitions and theorems. Proofs have been added for the basic theories. However, I would like to apologize for the fact that, due to the lack of effort on the author's part, there are multiple sections lacking content. Of course, for those in a hurry, it is recommended to skip the proofs and only use mathematical theorems as reference.

The basis of this book was the study material from the course “Applicable Systems Special Theory” of the Department of Complex Systems Science, Nagoya University Graduate School of Information Science. It is something that was born as a result of repeated trial and error throughout this course. I would like to thank the students who mastered the lesson content and worked together with me. I would also like to give special thanks to the graduates and students based in my research group who tackled shape optimization problems with me.

Moreover, the mathematicians based at the Japan Society for Industrial and Applied Mathematics gave me a lot of instruction on deciding the core principles of this book. Its overview is introduced in the “Afterword” and I would like to express my thanks.

This book is a revised edition of a book of the same title written in Japanese published by Morikita Publishing. I thank the people concerned of the publisher who permitted publishing an English version of the book. In the Japanese book, the readability was improved, thanks to Ms. Saho Kamimura's detailed amendments. These ideas have been taken over in this book. The publication of the English edition was actually due to a fortunate encounter with Mr. Masayuki Nakamura at an academic meeting. He recommended to realize my wish for publication of this book and has continuously given me words of encouragement to write and publish the book. In this book, some sections relating especially to the existing issues of optimum solutions were added from the Japanese book after receiving anonymous reviewers' comments. In regard to these additional sections, Dr. Julius Fergy T. Rabago and Dr. Masayuki Aino, who were doctoral students at Nagoya University in 2019, reviewed the theories and gave me a lot of valuable comments. Additionally,

a correction of a signification error about the evaluation of shape Hessians given in Chap. 9 of the Japanese version is provided in this book. Moreover, evaluation methods for computing second-order derivatives of cost functions via Lagrange multiplier methods were added in this book as additional contents. These additions were found during the editing process of the book and are considered imperative in this English edition. I really appreciate the hard work of the reviewers toward this book and Mr. Nakamura's kind work.

In the translation to English from Japanese on the original book, Crimson Interactive cooperated in rough translation. After the author's rewriting, polishing was done by Prof. Elliott Ginder in early chapters and Dr. Julius Fergy T. Rabago in the remaining chapters. The final check was performed by Springer. The author appreciates their powerful support to cover a lack of author's language ability.

Finally, I would like to thank my wife Yoshiko for her daily support throughout the process of writing this book. It would not have been possible without her constant love and kindness.

Nagoya, Japan
June 2020

Hideyuki Azegami

Contents

1	Basics of Optimal Design	1
1.1	Optimal Design Problem for a Stepped One-Dimensional Linear Elastic Body	1
1.1.1	State Determination Problem	3
1.1.2	An Optimal Design Problem	8
1.1.3	Cross-Sectional Derivatives	10
1.1.4	The Substitution Method	12
1.1.5	The Direct Differentiation Method	13
1.1.6	The Adjoint Variable Method	15
1.1.7	Optimality Conditions	21
1.1.8	Numerical Example	23
1.2	Comparison of the Direct Differentiation Method and the Adjoint Variable Method	24
1.2.1	The Direct Differentiation Method	26
1.2.2	The Adjoint Variable Method	27
1.3	An Optimal Design Problem of a Branched One-Dimensional Stokes Flow Field	29
1.3.1	State Determination Problem	30
1.3.2	An Optimal Design Problem	33
1.3.3	Cross-Sectional Derivatives	34
1.3.4	Optimality Conditions	38
1.3.5	Numerical Example	39
1.4	Summary	40
1.5	Practice Problems	41
2	Basics of Optimization Theory	45
2.1	Definition of Optimization Problems	45
2.2	Classification of Optimization Problems	48
2.3	Existence of a Minimum Point	51

2.4	Differentiation and Convex Functions	53
2.4.1	Taylor's Theorem	54
2.4.2	Convex Functions	56
2.4.3	Exercises in Differentiation and Convex Functions	59
2.5	Unconstrained Optimization Problems	61
2.5.1	A Necessary Condition for Local Minimizers	62
2.5.2	Sufficient Conditions for Local Minimizers	63
2.5.3	Sufficient Conditions for Global Minimizers	64
2.5.4	Example of Unconstrained Optimization Problem	64
2.5.5	Considerations Relating to the Solutions of Unconstrained Optimization Problems	65
2.6	Optimization Problems with Equality Constraints	65
2.6.1	A Necessary Condition for Local Minimizers	66
2.6.2	The Lagrange Multiplier Method	68
2.6.3	Sufficient Conditions for Local Minimizers	73
2.6.4	An Optimization Problem with an Equality Constraint ..	74
2.6.5	Direct Differentiation and Adjoint Variable Methods ..	76
2.6.6	Considerations Relating to the Solution of Optimization Problems with Equality Constraints	80
2.7	Optimization Problems Under Inequality Constraints	80
2.7.1	Necessary Conditions at Local Minimizers	82
2.7.2	Necessary and Sufficient Conditions for Global Minimizers	83
2.7.3	KKT Conditions	84
2.7.4	Sufficient Conditions for Local Minimizers	90
2.7.5	Sufficient Conditions for Global Minimizers Using the KKT Conditions	91
2.7.6	Example of an Optimization Problem Under an Inequality Constraint	92
2.7.7	Considerations Relating to the Solutions of Optimization Problems Under Inequality Constraints	94
2.8	Optimization Problems Under Equality and Inequality Constraints	94
2.8.1	The Lagrange Multiplier Method for Optimization Problems Under Equality and Inequality Constraints	96
2.8.2	Considerations Regarding Optimization Problems Under Equality and Inequality Constraints	97
2.9	Duality Theorem	98
2.9.1	Examples of the Duality Theorem	99
2.10	Summary	102
2.11	Practice Problems	103
3	Basics of Mathematical Programming	105
3.1	Problem Setting	105
3.2	Iterative Method	106

3.3	Gradient Method	107
3.4	Step Size Criterion	114
3.5	Newton Method	125
3.6	Augmented Function Methods	130
3.7	Gradient Method for Constrained Problems	132
3.7.1	Simple Algorithm	135
3.7.2	Complicated Algorithm	141
3.8	Newton Method for Constrained Problems	146
3.8.1	Simple Algorithm	149
3.8.2	Complicated Algorithm	157
3.9	Summary	157
3.10	Practice Problems	158
4	Basics of Variational Principles and Functional Analysis	159
4.1	Variational Principles	160
4.1.1	Hamilton's Principle	160
4.1.2	Minimum Principle of Potential Energy	162
4.1.3	Pontryagin's Minimum Principle	165
4.2	Abstract Spaces	171
4.2.1	Linear Space	171
4.2.2	Linear Subspaces	175
4.2.3	Metric Space	176
4.2.4	Normed Space	179
4.2.5	Inner Product Space	182
4.3	Function Spaces	183
4.3.1	Hölder Space	183
4.3.2	Lebesgue Space	185
4.3.3	Sobolev Space	188
4.3.4	Sobolev Embedding Theorem	193
4.4	Operators	196
4.4.1	Bounded Linear Operator	197
4.4.2	Trace Theorem	198
4.4.3	Calderón Extension Theorem	199
4.4.4	Bounded Bilinear Operators	200
4.4.5	Bounded Linear Functional	201
4.4.6	Dual Space	201
4.4.7	Rellich–Kondrachov Compact Embedding Theorem	207
4.4.8	Riesz Representation Theorem	208
4.5	Generalized Derivatives	209
4.5.1	Gâteaux Derivative	209
4.5.2	Fréchet Derivative	211
4.6	Function Spaces in Variational Principles	213
4.6.1	Hamilton's Principle	214
4.6.2	Minimum Principle of Potential Energy	216
4.6.3	Pontryagin's Minimum Principle	218

4.7	Summary	220
4.8	Practice Problems	220
5	Boundary Value Problems of Partial Differential Equations	223
5.1	Poisson Problem	223
5.1.1	Extended Poisson Problem	226
5.2	Abstract Variational Problem	228
5.2.1	Lax–Milgram Theorem	229
5.2.2	Abstract Minimization Problem	233
5.3	Regularity of Solutions	234
5.3.1	Regularity of Given Functions	235
5.3.2	Regularity of Boundary	235
5.4	Linear Elastic Problem	240
5.4.1	Linear Strain	240
5.4.2	Cauchy Tensor	242
5.4.3	Constitutive Equation	243
5.4.4	Equilibrium Equations of Force	245
5.4.5	Weak Form	246
5.4.6	Existence of Solution	247
5.5	Stokes Problem	249
5.6	Abstract Saddle Point Variational Problem	251
5.6.1	Existence Theorem of Solution	252
5.6.2	Abstract Saddle Point Problem	254
5.7	Summary	255
5.8	Practice Problems	256
6	Fundamentals of Numerical Analysis	259
6.1	Galerkin Method	260
6.1.1	One-Dimensional Poisson Problem	260
6.1.2	d -Dimensional Poisson Problem	266
6.1.3	Ritz Method	270
6.1.4	Basic Error Estimation	271
6.2	One-Dimensional Finite Element Method	272
6.2.1	Approximate Functions in Galerkin Method	273
6.2.2	Approximate Functions in Finite Element Method	275
6.2.3	Discretized Equations	278
6.2.4	Exercise Problem	282
6.3	Two-Dimensional Finite Element Method	286
6.3.1	Approximate Functions in Galerkin Method	286
6.3.2	Approximate Functions in Finite Element Method	288
6.3.3	Discretized Equations	290
6.3.4	Exercise Problem	295
6.4	Various Finite Elements	299
6.4.1	One-Dimensional Higher-Order Finite Elements	299
6.4.2	Triangular Higher-Order Finite Elements	303
6.4.3	Rectangular Finite Elements	306

6.4.4	Tetrahedral Finite Elements	309
6.4.5	Hexahedral Finite Elements	310
6.5	Isoparametric Finite Elements	310
6.5.1	Two-Dimensional Four-Node Isoparametric Finite Elements	312
6.5.2	Gaussian Quadrature	313
6.6	Error Estimation	318
6.6.1	Finite Element Division Sequence	319
6.6.2	Affine-Equivalent Finite Element Division Sequence	319
6.6.3	Interpolation Error Estimation	322
6.6.4	Error Estimation of Finite Element Solution	324
6.7	Summary	327
6.8	Practice Problems	327
7	Abstract Optimum Design Problem	331
7.1	Linear Spaces of Design Variables	332
7.2	State Determination Problem	333
7.3	Abstract Optimum Design Problem	334
7.4	Existence of an Optimum Solution	335
7.5	Derivatives of Cost Functions	338
7.5.1	Adjoint Variable Method	341
7.5.2	Lagrange Multiplier Method	342
7.5.3	Second-Order Fréchet Derivatives of Cost Functions	343
7.5.4	Second-Order Fréchet Derivative of Cost Function Using Lagrange Multiplier Method	345
7.6	Descent Directions of Cost Functions	346
7.6.1	Abstract Gradient Method	346
7.6.2	Abstract Newton Method	348
7.7	Solution of Abstract Optimum Design Problem	349
7.7.1	Gradient Method for Constrained Problems	350
7.7.2	Newton Method for Constrained Problems	354
7.8	Summary	356
8	Topology Optimization Problems of Density Variation Type	359
8.1	Set of Design Variables	363
8.2	State Determination Problem	365
8.3	Topology Optimization Problem of θ -Type	368
8.4	Existence of an Optimum Solution	370
8.5	Derivatives of Cost Functions	374
8.5.1	θ -Derivatives of Cost Functions	374
8.5.2	Second-Order θ -Derivative of Cost Functions	377
8.5.3	Second Order θ -Derivative of Cost Function Using Lagrange Multiplier Method	380
8.6	Descent Directions of Cost Functions	383
8.6.1	H^1 Gradient Method	383
8.6.2	H^1 Newton Method	387

8.7	Solution of Topology Optimization Problem of θ -Type.....	388
8.7.1	Gradient Method for Constrained Problems	389
8.7.2	Newton Method for Constrained Problems	390
8.8	Error Estimation.....	391
8.9	Topology Optimization Problem of Linear Elastic Body	397
8.9.1	State Determination Problem	398
8.9.2	Mean Compliance Minimization Problem	399
8.9.3	θ -Derivatives of Cost Functions	400
8.9.4	Second-Order θ -Derivatives of Cost Functions	402
8.9.5	Second-Order θ -Derivative of Cost Function Using Lagrange Multiplier Method	404
8.9.6	Numerical Example	406
8.10	Topology Optimization Problem of Stokes Flow Field	409
8.10.1	State Determination Problem	410
8.10.2	Mean Flow Resistant Minimization Problem	412
8.10.3	θ -Derivatives of Cost Functions	413
8.10.4	Second-Order θ -Derivatives of Cost Functions	415
8.10.5	Second-Order θ -Derivative of Cost Function Using Lagrange Multiplier Method	418
8.10.6	Numerical Example	419
8.11	Summary	423
8.12	Practice Problems	425
9	Shape Optimization Problems of Domain Variation Type	427
9.1	Set of Domain Variations and Definition of Shape Derivatives	431
9.1.1	Initial Domain	431
9.1.2	Sets of Domain Variations	433
9.1.3	Definitions of Shape Derivatives	435
9.2	Shape Derivatives of Jacobi Determinants	440
9.2.1	Shape Derivatives of Domain Jacobi Determinant and Domain Jacobi Inverse Matrix	441
9.2.2	Shape Derivatives of Boundary Jacobi Determinant and the Normal	442
9.3	Shape Derivatives of Functionals.....	446
9.3.1	Formulae Using Shape Derivative of a Function	446
9.3.2	Formulae Using Partial Shape Derivative of a Function	455
9.4	Variation Rules of Functions	461
9.5	State Determination Problem	464
9.6	Shape Optimization Problem of Domain Variation Type	467
9.7	Existence of an Optimum Solution.....	470
9.8	Derivatives of Cost Functions	477
9.8.1	Shape Derivative of f_i Using Formulae Based on Shape Derivative of a Function	477
9.8.2	Second-Order Shape Derivative of f_i Using Formulae Based on Shape Derivative of a Function	482

9.8.3	Second-Order Shape Derivative of Cost Function Using Lagrange Multiplier Method	486
9.8.4	Shape Derivative of f_i Using Formulae Based on Partial Shape Derivative of a Function	490
9.9	Descent Directions of Cost Functions.....	493
9.9.1	H^1 Gradient Method	494
9.9.2	H^1 Newton Method	501
9.10	Solution to Shape Optimization Problem of Domain Variation Type.....	502
9.10.1	Gradient Method for Constrained Problems	502
9.10.2	Newton Method for Constrained Problems	503
9.11	Error Estimation.....	504
9.12	Shape Optimization Problem of Linear Elastic Body	517
9.12.1	State Determination Problem	518
9.12.2	Mean Compliance Minimization Problem	519
9.12.3	Shape Derivatives of Cost Functions	520
9.12.4	Relation with Optimal Design Problem of Stepped One-Dimensional Linear Elastic Body	533
9.12.5	Numerical Example	537
9.13	Shape Optimization Problem of Stokes Flow Field	541
9.13.1	State Determination Problem	542
9.13.2	Mean Flow Resistance Minimization Problem	544
9.13.3	Shape Derivatives of Cost Functions	544
9.13.4	Relationship with Optimal Design Problem of One-Dimensional Branched Stokes Flow Field.....	556
9.13.5	Numerical Example	558
9.14	Summary	562
9.15	Practice Problems	564
Appendices	567
A.1	Basic Terminology	567
A.1.1	Open Sets, Closed Sets and Bounded Sets	567
A.1.2	Continuity of Functions	567
A.2	Positive Definiteness of Real Symmetric Matrix	569
A.3	Null Space, Image space and Farkas's Lemma	570
A.4	Implicit Function Theorem	572
A.5	Lipschitz Domain	573
A.6	Heat Conduction Problem	577
A.6.1	One-Dimensional Problem.....	577
A.6.2	d -Dimensional Problem.....	580
A.7	Classification of Second-Order Partial Differential Equations.....	583
A.8	Divergence Theorems	585

A.9	Inequalities	586
A.10	Ascoli–Arzelà Theorem	588
Answers to Practice Problems		591
Afterword		631
References		633
Index		641

Notation

Use of Letters

The use of variables will adhere to the following basic rules:

a, α, \dots	Lowercase italic Latin and Greek letters (Table 1) represent scalars, vectors, and functions
a, α, \dots	Bold lowercase italic Latin and Greek letters represent finite-dimensional vectors and functions with such ranges
$A, \mathcal{A}, \Gamma, \dots$	Capital italic Latin and Greek characters, as well as their cursive representations, represent sets
$\mathbf{A}, \Gamma, \dots$	Bold capital italic Latin and Greek letters represent finite dimension matrices and functions with such ranges
\mathcal{L}, \mathcal{H}	Represent Lagrange and Hamilton functions, respectively
a_A, a_{div}	Subscripts on upright Latin letters represent the initial of a terminology or an abbreviation

Below, m , n , and d denote natural numbers.

Sets

$\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$	Represent the set of natural numbers (positive integers), the integers, the rational numbers, the real numbers, and the complex numbers, respectively
\mathbb{R}^d	Represents the d -dimensional real linear space (real vector space)
$A = \{a_1, \dots, a_m\}$	Represents a set A consisting of a finite number of elements, a_1, \dots, a_m
$ A $	Represents the number of elements in a finite set A
$a \in A$	Indicates that a is an element of the set A

Table 1 Greek letters

Capital	Lowercase	Pronunciation	Capital	Lowercase	Pronunciation
A	α	alpha	N	ν	nu
B	β	beta	Ξ	ξ	xi
Γ	γ	gamma	O	o	omicron
Δ	δ	delta	Π	π	pi
E	ϵ	epsilon	P	ρ	rho
Z	ζ	zeta	Σ	σ	sigma
H	η	eta	T	τ	tau
Θ	θ	theta	Υ	υ	upsilon
I	ι	iota	Φ	ϕ	phi
K	κ	kappa	X	χ	chi
Λ	λ	lambda	Ψ	ψ	psi
M	μ	mu	Ω	ω	omega

$\{0\}$	Represents a set consisting of the single element 0
$\{a_k\}_{k \in \mathbb{N}}$	Denotes an infinite sequence, $\{a_1, a_2, \dots\}$
$A \subset B$	Indicates that the set A is a subset of the set B
$A \cup B, A \cap B, A \setminus C$	Represent the union, intersection, and subtraction operations of sets, respectively
$(0, 1), [0, 1], (0, 1]$	Denote the intervals $\{x \in \mathbb{R} \mid 0 < x < 1\}$, $\{x \in \mathbb{R} \mid 0 \leq x \leq 1\}$, $\{x \in \mathbb{R} \mid 0 < x \leq 1\}$, respectively

Vectors and Matrices

The following notation concerns vectors and matrices of \mathbb{R}^d , \mathbb{R}^m , and \mathbb{R}^n :

$\mathbf{x} = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$	Represents a d -dimensional vertical real vector. x_i represents the i -th element of \mathbf{x} . \mathbf{x}^\top represents the transposition of \mathbf{x}
$\mathbf{A} = (a_{ij})_{ij} \in \mathbb{R}^{m \times n}$	Represents the real matrix with m rows and n columns.
	Can also be written as $\mathbf{A} = (a_{ij})_{(i,j) \in \{1, \dots, m\} \times \{1, \dots, n\}}$
$\mathbf{0}_{\mathbb{R}^d}, \mathbf{0}_{\mathbb{R}^{m \times n}}$	Represents the zero elements of \mathbb{R}^d and $\mathbb{R}^{m \times n}$
$\mathbf{x} \geq \mathbf{0}_{\mathbb{R}^d}$	Represents $x_i \geq 0$ for $i \in \{1, \dots, d\}$
$\ \mathbf{x}\ _{\mathbb{R}^d, p}$	When $p \in [1, \infty)$, it represents the p -powered norm $\sqrt[p]{ x_1 ^p + \dots + x_d ^p}$ of $\mathbf{x} \in \mathbb{R}^d$. When $p = \infty$, it represents the maximum norm $\max\{ x_1 ^p, \dots, x_d ^p\}$. If there is no confusion, it is written as $\ \mathbf{x}\ _p$
$\mathbf{a} \cdot \mathbf{b}, \mathbf{A} \cdot \mathbf{B}$	Represents the inner product (scalar product) $\sum_{i \in \{1, \dots, m\}} a_i b_i$, $\sum_{(i,j) \in \{1, \dots, m\}^2} a_{ij} b_{ij}$ with respect to $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$ and $\mathbf{A} = (a_{ij})_{ij}$, $\mathbf{B} = (b_{ij})_{ij} \in \mathbb{R}^{m \times m}$

$\ \mathbf{x}\ _{\mathbb{R}^d}$	Represents the Euclidean norm $\sqrt{\mathbf{x} \cdot \mathbf{x}}$ of $\mathbf{x} \in \mathbb{R}^d$. If there is no confusion, it is written as $\ \mathbf{x}\ $
δ_{ij}	Represents the Kronecker delta $\delta_{ij} = 1$ ($i = j$), $\delta_{ij} = 0$ ($i \neq j$)
$\mathbf{I}_{\mathbb{R}^{m \times m}}$	Represents the unit matrix $(\delta_{ij})_{ij} \in \mathbb{R}^{m \times m}$. If there is no confusion, it is written as \mathbf{I}
$(\cdot)^s$	Represents $((\cdot)^\top + (\cdot)) / 2$

Domains and Functions

Here, functions on domains in \mathbb{R}^d are considered.

$\Omega \subset \mathbb{R}^d$	Represents a domain (simply connected open set) of \mathbb{R}^d
$\bar{\Omega}$	Represents the closure of Ω
$\partial\Omega$	Represents the boundary of Ω , i.e. $\bar{\Omega} \setminus \Omega$
$ \Omega $	Represents $\int_{\Omega} d\mathbf{x}$
S°	Represents the interior of a closed set S
\mathbf{v}	Represents an outward unit normal defined at the boundary $\partial\Omega$
$\boldsymbol{\tau}_1, \dots, \boldsymbol{\tau}_{d-1}$	Represents the tangent defined at the boundary $\partial\Omega$
κ	Represents $\nabla \cdot \mathbf{v}$ ($d-1$ times the mean curvature) defined at the boundary $\partial\Omega$
∇u	Represents the gradient $\partial u / \partial \mathbf{x} \in \mathbb{R}^d$ of a function $u : \mathbb{R}^d \rightarrow \mathbb{R}$
Δu	Represents the Laplace operator $\Delta = \nabla \cdot \nabla$ of a function $u : \mathbb{R}^d \rightarrow \mathbb{R}$
$\partial_\nu u$	Represents $(\mathbf{v} \cdot \nabla) u$ defined for a function $u : \Omega \rightarrow \mathbb{R}$ at the boundary $\partial\Omega$
$\partial_\nu \mathbf{u}$	Represents $(\mathbf{v} \cdot \nabla) \mathbf{u} = (\nabla \mathbf{u}^\top)^\top \mathbf{v}$ for function $\mathbf{u} : \Omega \rightarrow \mathbb{R}^d$ defined at the boundary $\partial\Omega$
$d\mathbf{x}, d\gamma, d\zeta$	Represents the measures used in the integration in the domain $\Omega \subset \mathbb{R}^d$, integration in the boundary $\Gamma \subset \partial\Omega$, and integration in the boundary of boundary $\partial\Gamma$
$\text{ess sup}_{\text{a.e. } \mathbf{x} \in \Omega} u(\mathbf{x}) $	Represents the essential bound of $u : \Omega \rightarrow \mathbb{R}$. The letters a.e. mean that it is almost everywhere on the measurable sets
χ_Ω	Represents the characteristic function $\chi_\Omega : \mathbb{R}^d \rightarrow \mathbb{R}$ ($\chi_\Omega(\Omega) = 1$, $\chi_\Omega(\mathbb{R}^d \setminus \bar{\Omega}) = 0$) with respect to a domain $\Omega \subset \mathbb{R}^d$

Banach Spaces

Here, V will be a normed space and X and Y are Banach spaces.

$\ x\ _V$	Represents the norm of $x \in V$. If there is no confusion, it is written as $\ x\ $
$f : X \rightarrow Y$	Represents a mapping (operator) from X to Y
$f(x) : X \ni x \mapsto f \in Y$	Represents a mapping expressing elements
$f \circ g$	Represents the composite mapping $f(g)$
$\mathcal{L}(X; Y)$	Represents the universal set of bounded linear operators from X to Y
$\mathcal{L}^2(X \times X; Y)$	Represents $\mathcal{L}(X; \mathcal{L}(X; Y))$
X'	Represents the dual space of X , in other words the universal set $\mathcal{L}(X; \mathbb{R})$ of bounded linear functions on X
$\langle y, x \rangle_{X' \times X}$	Represents the dual product of $x \in X$ and $y \in X'$. If there is no confusion, it is written as $\langle y, x \rangle$
$f'(x)[y]$	Represents the Fréchet derivative $\langle f'(x), y \rangle_{X' \times X}$ of $f : X \rightarrow \mathbb{R}$ at $x \in X$ with respect to an arbitrary variation $y \in X$
$f_x(x, y)[z], \partial_x f(x, y)[z]$	Represents the Fréchet partial derivative $\langle \partial f(x, y) / \partial x, z \rangle_{X' \times X}$ of $f : X \times Y \rightarrow \mathbb{R}$ at $(x, y) \in X \times Y$ with respect to an arbitrary variation $z \in X$. $f(x, y) / \partial x \in X'$ is written as $f_x(x, y)$
$C^k(X; Y)$	Represents the universal set of the $k \in \{0, 1, \dots\}$ -th Fréchet differentiable mappings
$C_{S'}^k(X; Y)$	Represents the universal set of the $k \in \{0, 1, \dots\}$ -th shape differentiable mappings
$C_{S'}^k(X; Y)$	Represents the universal set of the $k \in \{0, 1, \dots\}$ -th shape differentiable mappings
$C_{S^*}^k(X; Y)$	Represents the universal set of the $k \in \{0, 1, \dots\}$ -th partial shape differentiable mappings

Function Spaces

Here, Ω is a domain in \mathbb{R}^d .

$C(\Omega; \mathbb{R}^n), C^0(\Omega; \mathbb{R}^n)$	Represents the universal set of the continuous function $f : \Omega \rightarrow \mathbb{R}^n$ defined on Ω
$C_B(\Omega; \mathbb{R}^n), C_B^0(\Omega; \mathbb{R}^n)$	Represents the universal set of bounded functions in $C(\Omega; \mathbb{R}^n)$
$C_0(\Omega; \mathbb{R}^n)$	Represents the universal set of $f \in C(\Omega; \mathbb{R}^n)$ such that the support of f becomes a compact set of Ω

$C^k(\Omega; \mathbb{R}^n)$	Represents the universal set of $f \in C(\Omega; \mathbb{R}^n)$ for which up to the $k \in \{0, 1, \dots\}$ -th derivative of f belongs to $C(\Omega; \mathbb{R}^n)$
$C_B^k(\Omega; \mathbb{R}^n)$	Represents the universal set of $f \in C^k(\Omega; \mathbb{R}^n)$ for which up to the k -th derivative of f belongs to $C_B(\Omega; \mathbb{R}^n)$
$C_0^k(\Omega; \mathbb{R}^n)$	Represents $C^k(\Omega; \mathbb{R}^n) \cap C_0(\Omega; \mathbb{R}^n)$
$C^{k,\sigma}(\Omega; \mathbb{R}^n)$	Represents the universal set of $f \in C^k(\Omega; \mathbb{R}^n)$ for which up to the k -th derivative of f is a Hölder continuous function with the Hölder index $\sigma \in (0, 1]$. When $k = 0$ and $\sigma = 1$, the function is said to be Lipschitz continuous
$L^p(\Omega; \mathbb{R}^n)$	Represents the universal set of p -th powered Lebesgue integrable functions $f : \Omega \rightarrow \mathbb{R}^n$ for $p \in [1, \infty)$, and the universal set of functions which are essentially bounded for $p = \infty$
$W^{k,p}(\Omega; \mathbb{R}^n)$	Represents the universal set of functions for which up to the $k \in \{0, 1, \dots\}$ -th derivative belongs to $L^p(\Omega; \mathbb{R}^n)$
$W_0^{k,p}(\Omega; \mathbb{R}^n)$	Represents the closure of $C_0^\infty(\Omega; \mathbb{R})$ in $W^{k,p}(\Omega; \mathbb{R})$
$H^k(\Omega; \mathbb{R}^n)$	Represents $W^{k,2}(\Omega; \mathbb{R}^n)$
$H_0^k(\Omega; \mathbb{R}^n)$	Represents $W_0^{k,2}(\Omega; \mathbb{R}^n)$

Chapter 1

Basics of Optimal Design



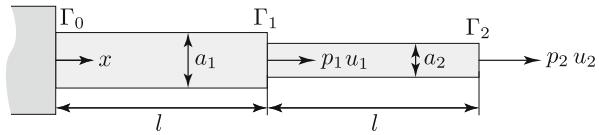
The main topic of this book is optimal design. In order to understand the mathematical structures involved in our study, we will begin by examining two simple problems. Upon finishing this book, the reader should be able to understand that even shape optimization problems of continuum structures possess the same formulation as the problems dealt with in this chapter. Moreover, even if the target continuum is changed from linear elastic body or flow field dealt in this book, the reader will recognize that their corresponding shape optimization problems maintain the fundamental structures introduced in this book.

Linear elastic solids and the Stokes flow field constitute the continuum used in our applications. We will construct optimal design problems related to one-dimensional linear elastic bodies and the one-dimensional Stokes flow field. We also show how to obtain optimality conditions for these problems. The conditions that we obtain will again be encountered in Chap. 9, where we will treat shape optimization problems of domain variation type for linear elastic bodies and the Stokes flow field, in two and three dimensions.

1.1 Optimal Design Problem for a Stepped One-Dimensional Linear Elastic Body

In order to understand the structure of optimal design problems, let us consider a mechanical system consisting of a one-dimensional linear elastic body with two cross-sectional areas, such as is shown in Fig. 1.1. The reason we refer to this system as one-dimensional is because, although in reality it is a three-dimensional body having a cross-sectional area and length, the x coordinate is taken in the length direction and, as is shown later, the displacement of the elastic body can be described as a function of x . In other words, it is assumed that the displacement

Fig. 1.1 1D linear elastic body with two cross-sectional areas



is given as a function on a one-dimensional vector space.¹ Furthermore, as will also be shown later, the linearity of the elastic body arises from the fact that, under the assumptions that the constitutive law is given by Hooke's law and that the deformation is infinitesimal, the outer force is a linear function of displacement.

Let us now define several constants and variables in detail. In particular, l is a constant representing length, a_1 and a_2 are cross-sectional areas, and $\mathbf{a} = (a_1, a_2)^\top \in \mathbb{R}^2$ is a vector with two components. In this book, \mathbb{R} denotes the set of all real numbers and $(\cdot)^\top$ represents the transpose. Moreover, bold lower-case Latin and Greek letters will be used in mathematical equations to represent finite-dimensional vectors. We remark that there exist positive constants a_{01} and a_{02} satisfying $a_i \geq a_{0i}$ for $i \in \{1, 2\}$. This can be expressed as $\mathbf{a} \geq \mathbf{a}_0$, where $\mathbf{a}_0 = (a_{01}, a_{02})^\top \in \mathbb{R}^2$. Similarly, letting p_1 and p_2 denote external forces acting on cross-sections Γ_1 and Γ_2 , and u_1 and u_2 be the corresponding displacements, we write $\mathbf{p} = (p_1, p_2)^\top \in \mathbb{R}^2$ and $\mathbf{u} = (u_1, u_2)^\top \in \mathbb{R}^2$.

Now let us consider an optimal design problem, where l and \mathbf{p} are assumed to be given. We treat \mathbf{a} as the design variable, due to the fact that once the cross-section \mathbf{a} is determined, the system we are attempting to design is uniquely determined. When a system is specified by determining \mathbf{a} , the variable \mathbf{u} which satisfies the system's state equation is called a state variable. In this book, the problem of finding the state variable is referred to as the state determination problem. The state determination problem that we are currently considering will be examined in detail in Sect. 1.1.1.

When the design variable \mathbf{a} and the state variable \mathbf{u} are given, we define real-valued functions of \mathbf{a} and \mathbf{u} representing the performance of the system. Such functions are called cost functions. In Sect. 1.1.2, considering that our current system is a structure supporting an external force, a function for measuring deformation and a function for imparting a volume constraint are chosen as the cost functions. The cost functions are then used to formulate the optimal design problem through defining objective and constraint functions.

The condition which holds when an optimal solution is used in an optimal design problem constructed in this manner is called an optimality condition. An optimality condition for the current one-dimensional elastic body problem is presented in Sect. 1.1.7. For this reason, the derivative of the cost function with respect to the variation of a design variable is defined in Sect. 1.1.3, and ways to obtain them are considered from Sects. 1.1.4 to 1.1.6. These results should perhaps be presented after an explanation has been given regarding a main theorem of optimization theory,

¹The finite-dimensional vector space considered here can also be called a Euclid space. Moreover, vector space is synonymous with linear space (see Definition 4.2.1).

which is given in Chap. 2. Nevertheless, this book sets a priority on obtaining a practical understanding of how to make use of this theorem in optimization theory.

1.1.1 State Determination Problem

Let us go through the process involved in constructing an optimal design problem. We will begin by defining a mechanical system which is the target of the design. When the design variables are specified, this system reverts to a mechanical problem constructed by standard equations of motion and boundary conditions. We refer to this problem as a state determination problem, and we examine its construction based on mechanical principles. Readers who are knowledgeable in the field of mechanics are invited to skip this section.

Before analyzing our one-dimensional elastic body, we review the fact that the equilibrium equation of forces can be obtained from minimality conditions of a potential energy [171]. The next exercise concerns the definition of potential energy.

Exercise 1.1.1 (Potential Energy of a Simple Spring System) Consider a spring system with a single degree of freedom, such as is shown in Fig. 1.2. Let k and p denote positive numbers representing the spring constant and an external force, respectively. Moreover, let the external force be conservative, that is a constant force generated anywhere on \mathbb{R} and $u \in \mathbb{R}$ be the displacement when the spring and the external force are in balance. Assume that

$$ku - p = 0$$

holds. Find the potential energy of the spring system when $u = 0$ is set as the point of reference. \square

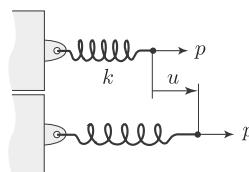


Fig. 1.2 A spring system with a single degree of freedom. The top figure represents the system's initial state, and the bottom figure illustrates the balanced state of forces

Answer In mechanics, potential energy is defined as an amount of energy which expresses the capacity to do work. When $u = 0$ is the point of reference, the potential energy is obtained by integrating the unbalanced force $kv - p$ (v denotes an intermediate displacement) over a displacement from 0 to u :

$$\pi(u) = \int_0^u (kv - p) \, dv = \frac{1}{2}ku^2 - pu. \quad (1.1.1)$$

□

The first and the second terms on the right-hand side of Eq. (1.1.1) are called the internal potential energy and the external potential energy, respectively. Notice that the internal potential energy is the part which acquires the ability to do work (potential) and is therefore positive. On the other hand, the external potential energy is the part which has already done work (the directions of the force and the displacement are the same), and so it is negative. Here we remark that, although potential energy is related to the stored energy (Hamiltonian) which appears in the law of conservation of energy (see Practice 4.3), the two are in fact different entities.

If a potential energy of π is obtained, the force equilibrium equation is given by the stationary condition of the potential energy:

$$\frac{d\pi}{du} = ku - p = 0.$$

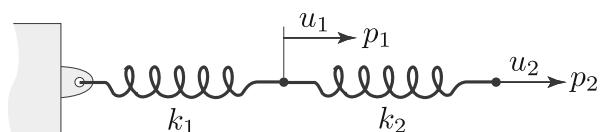
The fact that the potential energy is minimized at this point is a consequence of the following:

$$\frac{d^2\pi}{du^2} = k > 0.$$

Once the notion of potential energy is understood, one can also assess the potential energy of spring systems with two degrees of freedom. As in the following exercise, the same idea can be applied to a force equilibrium equation of a two-degree-of-freedom spring system.

Exercise 1.1.2 (Potential Energy in a 2DOF Spring System) Consider a spring system consisting of two degrees of freedom, such as is shown in Fig. 1.3. Here, k_1 and k_2 are positive constants representing the spring constants, $\mathbf{p} = (p_1, p_2)^\top \in \mathbb{R}^2$ is a constant vector representing external forces, and $\mathbf{u} = (u_1, u_2)^\top \in \mathbb{R}^2$ denotes the displacement when in a balanced state with \mathbf{p} . In this case, obtain the potential energy when $\mathbf{u} = \mathbf{0}_{\mathbb{R}^2}$ ($\mathbf{0}_{\mathbb{R}^2}$ denotes $(0, 0)^\top$ in this book) is the point of reference.

Fig. 1.3 A two-degree-of-freedom spring system



Also, find the force equilibrium equation using the stationary condition of the potential energy. \square

Answer The potential energy of the system can be obtained by adding together the internal and external potential energies:

$$\pi(\mathbf{u}) = \frac{1}{2}k_1u_1^2 + \frac{1}{2}k_2(u_2 - u_1)^2 - (p_1u_1 + p_2u_2).$$

One has the stationarity condition of potential energy:

$$\begin{aligned}\frac{\partial\pi}{\partial u_1} &= k_1u_1 - k_2(u_2 - u_1) - p_1 = 0, \\ \frac{\partial\pi}{\partial u_2} &= k_2(u_2 - u_1) - p_2 = 0,\end{aligned}$$

which can be used to obtain the force equilibrium equation. These equations can be written as

$$\begin{pmatrix} k_1 + k_2 & -k_2 \\ -k_2 & k_2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}. \quad (1.1.2)$$

\square

The fact that a vector \mathbf{u} satisfying the stationary condition Eq. (1.1.2) minimizes the potential energy π can be shown using the approach of Exercise 2.5.8. This fact will be omitted for now.

We have confirmed that the force equilibrium equation can be obtained via minimality conditions of the potential energy, so now let us apply these conditions to the one-dimensional linear elastic body shown in Fig. 1.1. First, similar to Exercise 1.1.2, the external potential energy can be given as

$$\pi_E(\mathbf{u}) = -\mathbf{p} \cdot \mathbf{u}. \quad (1.1.3)$$

In this book, $\mathbf{p} \cdot \mathbf{u} = \mathbf{p}^\top \mathbf{u}$ represents the inner product of a finite dimensional vector space.

Next, let us find the internal potential energy. Let $x \in \mathbb{R}$ be the coordinate in the length direction, where the cross-section Γ_0 in Fig. 1.1 is taken as the origin. In this case, the displacement at $x \in [0, 2l]$ is assumed to be given by

$$u(x) = \begin{cases} u_1 \frac{x}{l} & x \in [0, l) \\ (u_2 - u_1) \frac{x}{l} + 2u_1 - u_2 & x \in [l, 2l] \end{cases} \quad (1.1.4)$$

from the linearized elasticity assumption explained in the beginning of Sect. 1.1. In other words, it is assumed that three-dimensional deformations are not considered. Here, $[0, 2l]$ represents the interval $\{x \in \mathbb{R} \mid 0 \leq x \leq 2l\}$.

We would now like to interrupt our discussion in order to explain principles regarding the representation of sets and functions used in this book. Sets are defined in the format $\{x \in \mathbb{R} \mid 0 \leq x \leq 2l\}$, where the linear space (defined in Chap. 4) or an underlying set is written in the position of \mathbb{R} . Conditions satisfied by elements of the set are written after the \mid symbol. In particular, $[0, l)$ represents the interval $\{x \in \mathbb{R} \mid 0 \leq x < l\}$ and $(0, l)$ represents $\{x \in \mathbb{R} \mid 0 < x < l\}$. In Eq. (1.1.4), $u(x)$ is continuous at $x = l$ and therefore $[0, l)$ of Eq. (1.1.4) can be written as $[0, l]$ or as $(0, l)$. Hence, in this book, the domain of definition of a function is defined to be an open set (see Sect. A.1.1), and the function's boundary values are defined using properties of continuity (called the trace of the function (Theorem 4.4.2)). On the other hand, taking Eq. (1.1.4) as an example, the domain and range of the function $u(x)$ are expressed using the notation $u : (0, 2l) \rightarrow \mathbb{R}$, where \rightarrow designates that the mapping is from the domain $(0, 2l)$ into the range of real numbers \mathbb{R} . When specifying elements, we will sometimes write $u(x) : (0, 2l) \ni x \mapsto u \in \mathbb{R}$. Becoming too caught up with the wording of functions or variables in this book may lead to confusion, because functions themselves become variables from Chap. 4 onwards. Therefore, let us remember that the mapping notation is important in such cases.

Let us now return to our original discussion. In mechanics, equations relating variables representing phenomena such as force and displacement, or temperature and heat are called constitutive equations or constitutive laws. Hooke's law is used in the case of linear elastic bodies. Hooke's law relates the strain of a material (its rate of deformation)

$$\varepsilon(u) = \frac{du}{dx} \quad (1.1.5)$$

with its stress $\sigma(u)$ (force acting per unit area) via

$$\sigma(u) = e_Y \varepsilon(u). \quad (1.1.6)$$

Here, e_Y is assumed to be given by a material-specific positive constant called the modulus of longitudinal elasticity, or Young's modulus. In the one-dimensional linear elastic body of Fig. 1.1, it may be assumed that e_Y is given by a discontinuous function such as $e_Y : (0, 2l) \rightarrow \mathbb{R}$, but for the sake of simplicity we shall assume that it is given by a positive real constant. Furthermore, the mechanical quantity defined using the stress and the strain:

$$w(u) = \frac{1}{2} \sigma(u) \varepsilon(u) \quad (1.1.7)$$

is called the strain energy density (internal potential energy density or elastic potential energy density). The fact that w is an energy per unit volume can also be confirmed from the fact that its units are $[\text{N m/m}^3]$ in the international system of units (SI). Using these definitions, the internal potential energy of the one-

dimensional linear elastic body in Fig. 1.1 is given by

$$\pi_I(\mathbf{u}) = \int_0^l w(u) a_1 dx + \int_l^{2l} w(u) a_2 dx. \quad (1.1.8)$$

Since the internal and external potential energies of the one-dimensional elastic body in Fig. 1.1 were obtained using Eqs. (1.1.8) and (1.1.3), the total potential energy with $\mathbf{u} = \mathbf{0}_{\mathbb{R}^2}$ as a reference point is given by

$$\begin{aligned} \pi(\mathbf{u}) &= \pi_I(\mathbf{u}) + \pi_E(\mathbf{u}) \\ &= \frac{1}{2} \frac{e_Y}{l} a_1 u_1^2 + \frac{1}{2} \frac{e_Y}{l} a_2 (u_2 - u_1)^2 - p_1 u_1 - p_2 u_2. \end{aligned} \quad (1.1.9)$$

Therefore, the stationary condition of π is expressed as

$$\begin{aligned} \frac{\partial \pi}{\partial u_1} &= \frac{e_Y}{l} a_1 u_1 - \frac{e_Y}{l} a_2 (u_2 - u_1) - p_1 = 0, \\ \frac{\partial \pi}{\partial u_2} &= \frac{e_Y}{l} a_2 (u_2 - u_1) - p_2 = 0, \end{aligned}$$

which can also be written as

$$\frac{e_Y}{l} \begin{pmatrix} a_1 + a_2 & -a_2 \\ -a_2 & a_2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}. \quad (1.1.10)$$

This leads us to the problem of determining the displacement when external forces act on the one-dimensional linear elastic body in Fig. 1.1.

Problem 1.1.3 (Stepped 1D Linear Elastic Body) Let $l \in \mathbb{R}$, $e_Y \in \mathbb{R}$, $\mathbf{p} \in \mathbb{R}^2$ and $\mathbf{a} \in \mathbb{R}^2$ be given with respect to the one-dimensional linear elastic body of Fig. 1.1. Find the displacement $\mathbf{u} \in \mathbb{R}^2$ that satisfies

$$\mathbf{K}(\mathbf{a}) \mathbf{u} = \mathbf{p}, \quad (1.1.11)$$

where Eq. (1.1.11) of course represents Eq. (1.1.10). \square

In this book, matrices are expressed using bold capital Latin and Greek letters, such as \mathbf{K} .

Anticipating future developments, let us take a look at an alternative way of expressing Problem 1.1.3. With respect to Problem 1.1.3,

$$\mathcal{L}_S(\mathbf{a}, \mathbf{u}, \mathbf{v}) = \mathbf{v} \cdot (-\mathbf{K}(\mathbf{a}) \mathbf{u} + \mathbf{p}) \quad (1.1.12)$$

will be called a Lagrange function for a state determination problem (defined in Chap. 2). Here, $\mathbf{u} \in \mathbb{R}^2$ is not necessarily the solution of Problem 1.1.3 and $\mathbf{v} \in \mathbb{R}^2$ has been introduced as a Lagrange multiplier with respect to Eq. (1.1.11). The Lagrange multiplier with respect to a state equation is also referred to as an adjoint variable. Here, $\mathbf{u} \in \mathbb{R}^2$, which satisfies

$$\mathcal{L}_S(\mathbf{a}, \mathbf{u}, \mathbf{v}) = 0 \quad (1.1.13)$$

for all $\mathbf{v} \in \mathbb{R}^2$, has the same value as the solution of Problem 1.1.3. This is because if Eq. (1.1.11) is to be satisfied, then Eq. (1.1.13) holds for all $\mathbf{v} \in \mathbb{R}^2$. The converse also holds.

The condition under which Eq. (1.1.13) is satisfied for all $\mathbf{v} \in \mathbb{R}^2$ is called the principle of virtual work. The reason for this is that the potential energy

$$\pi(\mathbf{u}) = \frac{1}{2} \mathbf{u} \cdot (\mathbf{K}(\mathbf{a}) \mathbf{u}) - \mathbf{p} \cdot \mathbf{u}$$

has a stationary condition with respect to an arbitrary variation $d\mathbf{u} \in \mathbb{R}^2$ of \mathbf{u} (the virtual displacement), given by

$$d\pi(\mathbf{u}) = \mathcal{L}_S(\mathbf{a}, \mathbf{u}, d\mathbf{u}) = 0.$$

1.1.2 An Optimal Design Problem

Having defined the state determination problem, let us now use it to construct an optimal design problem. Let us first define the cost function. With respect to the solution \mathbf{u} of the state determination problem (Problem 1.1.3), the following quantity will be referred to as the mean compliance:

$$f_0(\mathbf{u}) = (p_1 \ p_2) \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \mathbf{p} \cdot \mathbf{u}. \quad (1.1.14)$$

Here, f_0 is equivalent to the mechanical quantity known as external work. Nevertheless, in Chaps. 8 and 9, a cost function measuring the ease of deformation (compliance) of linear elastic bodies will be given and used to define an extended notion of mean compliance. The naming in those cases is chosen in order to not imply work done by external forces. The fact that f_0 of Eq. (1.1.14) is a real-valued function representing an ease of deformation can be explained as follows. Since \mathbf{u} is a vector representing the ease of deformation, it is not simply a real number. Here, if f_0 is thought of as a function weighted by \mathbf{p} in order to convert \mathbf{u} into a real number, then it can be understood that f_0 is a real-valued function expressing the

ease of deformation. Relatedly,

$$f_1(\mathbf{a}) = l(a_1 + a_2) - c_1 = (l \ l) \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} - c_1 \quad (1.1.15)$$

will be referred to as a constraint function with respect to volume. Here, c_1 is a positive constant representing an upper bound on the volume. In this section, f_0 and f_1 are defined as cost functions for an optimal design problem. Throughout this book, cost functions will be denoted by f_0, f_1, \dots, f_m , where f_0 will denote the objective function, and f_1, \dots, f_m will denote constraint functions.

Let us define an optimal design problem with respect to the one-dimensional linear elastic body in Fig. 1.1 using the previous cost functions as in Problem 1.1.4. Hereinafter, linear spaces of the design variable \mathbf{a} and the state variable \mathbf{u} will be denoted as $X = \mathbb{R}^2$ and $U = \mathbb{R}^2$, respectively, and are called the linear space of design variables and the linear space of state variables. Moreover, with respect to a constant vector $\mathbf{a}_0 = (a_{01}, a_{02})^\top > \mathbf{0}_{\mathbb{R}^2}$, we have the admissible set of design variables:

$$\mathcal{D} = \{\mathbf{a} \in X \mid \mathbf{a} \geq \mathbf{a}_0\}. \quad (1.1.16)$$

In this book, capital Latin and Greek letters (including decorative scripts) are used for sets. Symbols relating to the sets X , U and \mathcal{D} will be used with a unified meaning, even in the setting of optimal design problems in function space (beginning in Chap. 7).

Problem 1.1.4 (Mean Compliance Minimization) Let $X = \mathbb{R}^2$, $U = \mathbb{R}^2$, and \mathcal{D} be given by Eq. (1.1.16). If $f_0(\mathbf{u})$ and $f_1(\mathbf{a})$ are given by Eqs. (1.1.14) and (1.1.15), respectively, find \mathbf{a} satisfying

$$\min_{(\mathbf{a}, \mathbf{u}) \in \mathcal{D} \times U} \{f_0(\mathbf{u}) \mid f_1(\mathbf{a}) \leq 0, \text{ Problem 1.1.3}\}. \quad \square$$

We remark that Problem 1.1.4 should probably be written as follows: find

$$\mathbf{a}^* = \arg \min_{\mathbf{a} \in \mathcal{D}} \{f_0(\mathbf{u}) \mid f_1(\mathbf{a}) \leq 0, \mathbf{u} \in U, \text{ Problem 1.1.3}\}.$$

Here, $\arg \min_{\mathbf{x} \in X} f(\mathbf{x})$ denotes a point \mathbf{x} in the domain X where f attains its minimum value. However, to simplify the expression, it will be written as shown in Problem 1.1.4. Moreover, simple constraints such as $\mathbf{a} \geq \mathbf{a}_0$ given in Eq. (1.1.16) will sometimes be referred to as side constraints. Later on, solutions satisfying optimality conditions will be sought while disregarding side constraints. A solution is then chosen from those candidates that satisfy the side constraints. Therefore, we assume that \mathbf{a} is an interior point of \mathcal{D} ($\mathbf{a} \in \mathcal{D}^\circ$) and when some of the side constraints are activated, we include them in the inequality constraints (Practice 1.4).

Problem 1.1.4 is an optimization problem with an equality and inequality constraint. Optimality conditions satisfied by solutions of this type of problem will be discussed in detail in Chap. 2 and methods for their numerical solutions will be considered in Chap. 3. An explanation regarding the details of these will be omitted at present, and we will look at how the optimality conditions are obtained by following a set of formal procedures. In the next section, a method for obtaining the derivatives of f_0 and f_1 with respect to variations of the design variable \mathbf{a} is considered. These results are used at once with the optimality conditions in Sect. 1.1.7.

Moreover, in the numerical solutions of optimum design problems, the calculation of derivatives of cost functions with respect to an arbitrary variation of design variable becomes the pivotal ingredient. In this case, the state determination problem, which will be a boundary value problem of partial differential equations in Chaps. 8 and 9, becomes an equality constraint. Then, an understanding of how to calculate the derivatives of cost functions f_0 in the next subsection will help the reader to understand the method used to obtain the derivatives of cost functions in Chaps. 8 and 9. In order to help the reader's understanding, we use the same notation as in Chaps. 8 and 9 as much as possible.

In the next subsection, we will obtain the derivative of cost function f_0 using three methods and confirm that those results accord. However, in Chaps. 8 and 9, we will use only one of them for convenience.

1.1.3 Cross-Sectional Derivatives

We call derivatives of f_0 and f_1 with respect to the variation of the cross-sectional area \mathbf{a} cross-sectional derivatives.

Let us start by considering the cross-sectional derivative of f_1 to which the usual definition of differentiation can be applied. Since f_1 is defined as a function of \mathbf{a} in Eq. (1.1.15), its partial derivative with respect to \mathbf{a} can be obtained as

$$f_{1\mathbf{a}} = \frac{\partial f_1}{\partial \mathbf{a}} = \begin{pmatrix} \partial f_1 / \partial a_1 \\ \partial f_1 / \partial a_2 \end{pmatrix} = \begin{pmatrix} l \\ l \end{pmatrix} = \mathbf{g}_1. \quad (1.1.17)$$

In this book, the partial derivative $\partial f_1 / \partial \mathbf{a}$ will be written as $f_{1\mathbf{a}}$. Here, note that $f_{1\mathbf{a}}$ is a column vector because, although f_1 is a real number, \mathbf{a} is a column vector. On the other hand, the Taylor expansion (Theorem 2.4.2) of f_1 at \mathbf{a} with respect to an arbitrary $\mathbf{b} \in X$ is

$$\begin{aligned} f_1(\mathbf{a} + \mathbf{b}) &= f_1(\mathbf{a}) + f'_1(\mathbf{a})[\mathbf{b}] + o(\|\mathbf{b}\|_{\mathbb{R}^2}) \\ &= f_1(\mathbf{a}) + \mathbf{g}_1 \cdot \mathbf{b} + o(\|\mathbf{b}\|_{\mathbb{R}^2}). \end{aligned} \quad (1.1.18)$$

Here $f'_1(\mathbf{a})[\mathbf{b}]$ represents the first-order variation of f_1 at \mathbf{a} with respect to the variation \mathbf{b} . Moreover, $o(\cdot)$ denotes the Bachmann–Landau little- o symbol defined by $\lim_{\epsilon \rightarrow 0} o(\epsilon)/\epsilon = 0$, where $\|\mathbf{b}\|_{\mathbb{R}^2} = \sqrt{|b_1|^2 + |b_2|^2}$. We also remark that $o(\|\mathbf{b}\|_{\mathbb{R}^2}) = 0$ with respect to f_1 . In view of Eq. (1.1.18), since $f'_1(\mathbf{a})[\mathbf{b}] = f_{1a} \cdot \mathbf{b} = \mathbf{g}_1 \cdot \mathbf{b}$, we note that $f'_1(\mathbf{a})[\mathbf{b}]$ is a linear function with respect to \mathbf{b} . In other words, the corresponding vector \mathbf{g}_1 of the inner product with respect to \mathbf{b} has been found. When the equation can be written as $f'_1(\mathbf{a})[\mathbf{b}] = \mathbf{g}_1 \cdot \mathbf{b}$ we say that it is differentiable and that $f'_1(\mathbf{a})[\mathbf{b}]$ is the cross-sectional derivative of f_1 at \mathbf{a} . We refer to $\mathbf{g}_1 \in \mathbb{R}^2$ as the cross-sectional-area gradient of f_1 .

Next, let us consider the cross-sectional derivative of f_0 in a similar way. Although f_0 is a function of \mathbf{u} as in Eq. (1.1.14), it is not explicitly a function of \mathbf{a} . However, \mathbf{u} is assumed to satisfy the state equation (Problem 1.1.3) for a given \mathbf{a} , so that \mathbf{u} varies with any variation of \mathbf{a} . In other words, \mathbf{u} is a function of \mathbf{a} . Let us now write

$$\tilde{f}_0(\mathbf{a}) = \{f_0(\mathbf{u}) \mid (\mathbf{a}, \mathbf{u}) \in \mathcal{D} \times U, \text{ Problem 1.1.3}\}, \quad (1.1.19)$$

and suppose that we have found a linear function $\tilde{f}'_0(\mathbf{a})[\mathbf{b}]$ with respect to \mathbf{b} satisfying

$$\tilde{f}_0(\mathbf{a} + \mathbf{b}) = \tilde{f}_0(\mathbf{a}) + \tilde{f}'_0(\mathbf{a})[\mathbf{b}] + o(\|\mathbf{b}\|_{\mathbb{R}^2}),$$

where we write $\tilde{f}'_0(\mathbf{a})[\mathbf{b}] = \mathbf{g}_0 \cdot \mathbf{b}$ for a certain $\mathbf{g}_0 \in \mathbb{R}^2$. Then f_0 is said to be differentiable with respect to \mathbf{a} , $\tilde{f}'_0(\mathbf{a})[\mathbf{b}]$ is called the cross-sectional derivative of f_0 at \mathbf{a} , and \mathbf{g}_0 is called the cross-sectional-area gradient.

Furthermore, with respect to a function $\mathbf{g}_0 : X \rightarrow \mathbb{R}^2$, whenever there exists $\mathbf{g}'_0(\mathbf{a})[\mathbf{b}_2]$ which is linear in \mathbf{b}_2 satisfying

$$\begin{aligned} \mathbf{g}_0(\mathbf{a} + \mathbf{b}_2) \cdot \mathbf{b}_1 &= \mathbf{g}_0(\mathbf{a}) \cdot \mathbf{b}_1 + \mathbf{g}'_0(\mathbf{a})[\mathbf{b}_2] \cdot \mathbf{b}_1 + o(\|\mathbf{b}_2\|_{\mathbb{R}^2}) \\ &= \mathbf{g}_0(\mathbf{a}) \cdot \mathbf{b}_1 + \tilde{f}''_0(\mathbf{a})[\mathbf{b}_1, \mathbf{b}_2] + o(\|\mathbf{b}_2\|_{\mathbb{R}^2}), \end{aligned}$$

which is expressible as $\mathbf{g}'_0(\mathbf{a})[\mathbf{b}_2] = \mathbf{H}_0 \mathbf{b}_2$ for a certain $\mathbf{H}_0 \in \mathbb{R}^{2 \times 2}$, then f_0 is second-order differentiable and $\mathbf{H}_0 \in \mathbb{R}^{2 \times 2}$ is referred to as the Hesse matrix or the Hessian (Definition 2.4.1) of f_0 at \mathbf{a} . This is equivalent to the condition that $\tilde{f}''_0(\mathbf{a})[\mathbf{b}_1, \mathbf{b}_2]$ is a bilinear function of \mathbf{b}_1 and \mathbf{b}_2 , and $\tilde{f}''_0(\mathbf{a})[\mathbf{b}_1, \mathbf{b}_2]$ is referred to as the second-order cross-sectional derivative. In this book, $\mathbb{R}^{m \times n}$ represents the set of all real matrices consisting of m rows and n columns.

Using these definitions, if $\tilde{f}_0(\mathbf{a})$ is second-order differentiable with respect to \mathbf{a} , the Taylor expansion (Theorem 2.4.2) of $\tilde{f}_0(\mathbf{a})$ at \mathbf{a} can be written

$$\tilde{f}_0(\mathbf{a} + \mathbf{b}) = \tilde{f}_0(\mathbf{a}) + \mathbf{g}_0 \cdot \mathbf{b} + \frac{1}{2} \tilde{f}''_0(\mathbf{a})[\mathbf{b}, \mathbf{b}] + o(\|\mathbf{b}\|_{\mathbb{R}^2}^2).$$

Later on, \mathbf{g}_0 is used under the condition that f_0 takes an extreme value. Moreover, $\tilde{f}_0''(\mathbf{a}) [\mathbf{b}, \mathbf{b}]$ (equivalently, \mathbf{H}_0) will be used in conditions to guarantee that minimum values are obtained. In the next section we will consider how to obtain \mathbf{g}_0 and \mathbf{H}_0 .

1.1.4 The Substitution Method

Let us now obtain \mathbf{g}_0 and \mathbf{H}_0 directly from $\tilde{f}_0(\mathbf{a})$ by direct substitution of the state equation into the cost function. We remark that this method cannot be used in more complicated problems. However, we shall use it here in order to verify results obtained from the direct differentiation and adjoint variable methods (shown later).

The solution of the state equation (Eq. (1.1.11)) is obtained as

$$\mathbf{u} = \mathbf{K}^{-1}(\mathbf{a}) \mathbf{p} = \frac{l}{e_Y} \begin{pmatrix} \frac{1}{a_1} & \frac{1}{a_1} \\ \frac{1}{a_1} & \frac{1}{a_1} + \frac{1}{a_2} \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \frac{l}{e_Y} \begin{pmatrix} \frac{p_1 + p_2}{a_1} \\ \frac{p_1 + p_2}{a_1} + \frac{p_2}{a_2} \end{pmatrix}. \quad (1.1.20)$$

Since $\tilde{f}_0(\mathbf{a})$ is defined by Eq. (1.1.19), the following equation can be obtained:

$$\tilde{f}_0(\mathbf{a}) = \mathbf{p} \cdot (\mathbf{K}^{-1}(\mathbf{a}) \mathbf{p}) = \frac{l}{e_Y} \left(\frac{(p_1 + p_2)^2}{a_1} + \frac{p_2^2}{a_2} \right), \quad (1.1.21)$$

from which we get

$$\mathbf{g}_0 = \begin{pmatrix} \frac{\partial \tilde{f}_0}{\partial a_1} \\ \frac{\partial \tilde{f}_0}{\partial a_2} \end{pmatrix} = \begin{pmatrix} \mathbf{p} \cdot \left(\frac{\partial \mathbf{K}^{-1}}{\partial a_1} \mathbf{p} \right) \\ \mathbf{p} \cdot \left(\frac{\partial \mathbf{K}^{-1}}{\partial a_2} \mathbf{p} \right) \end{pmatrix} = \frac{l}{e_Y} \begin{pmatrix} -\frac{(p_1 + p_2)^2}{a_1^2} \\ -\frac{p_2^2}{a_2^2} \end{pmatrix}. \quad (1.1.22)$$

Similarly, the Hesse matrix is expressed as

$$\mathbf{H}_0 = \begin{pmatrix} \frac{\partial^2 \tilde{f}_0}{\partial a_1 \partial a_1} & \frac{\partial^2 \tilde{f}_0}{\partial a_1 \partial a_2} \\ \frac{\partial^2 \tilde{f}_0}{\partial a_2 \partial a_1} & \frac{\partial^2 \tilde{f}_0}{\partial a_2 \partial a_2} \end{pmatrix} = \begin{pmatrix} \frac{\partial \mathbf{g}_0}{\partial a_1} & \frac{\partial \mathbf{g}_0}{\partial a_2} \end{pmatrix}$$

$$\begin{aligned}
&= \begin{pmatrix} \mathbf{p} \cdot \left(\frac{\partial^2 \mathbf{K}^{-1}}{\partial a_1 \partial a_1} \mathbf{p} \right) \mathbf{p} \cdot \left(\frac{\partial^2 \mathbf{K}^{-1}}{\partial a_1 \partial a_2} \mathbf{p} \right) \\ \mathbf{p} \cdot \left(\frac{\partial^2 \mathbf{K}^{-1}}{\partial a_1 \partial a_2} \mathbf{p} \right) \mathbf{p} \cdot \left(\frac{\partial^2 \mathbf{K}^{-1}}{\partial a_2 \partial a_2} \mathbf{p} \right) \end{pmatrix} \\
&= \frac{l}{e_Y} \begin{pmatrix} \frac{2(p_1 + p_2)^2}{a_1^3} & 0 \\ 0 & \frac{2p_2^2}{a_2^3} \end{pmatrix}. \tag{1.1.23}
\end{aligned}$$

Whenever $a_1, a_2 > 0$, the eigenvalues of \mathbf{H}_0 are positive and so \mathbf{H}_0 is positive definite (Definition 2.4.5). From this property, based on Theorem 2.4.6 (shown later), $\tilde{f}_0(\mathbf{a})$ can be shown to be a convex function (Definition 2.4.3). The convexity of $\tilde{f}_0(\mathbf{a})$ is used as a sufficient condition for showing minimality in Sect. 1.1.7.

1.1.5 The Direct Differentiation Method

Next, let us also obtain \mathbf{g}_0 and \mathbf{H}_0 via the direct differentiation method, which utilizes the chain rule of differentiation for composite functions. The details of the direct differentiation method are presented in Sect. 2.6.5.

Note that \mathbf{u} is determined with respect to \mathbf{a} such that Eq. (1.1.11) is satisfied. Thus, if \tilde{f}_0 of Eq. (1.1.19) is Taylor expanded around \mathbf{a} we have

$$\begin{aligned}
\tilde{f}_0(\mathbf{a} + \mathbf{b}) &= f_0(\mathbf{u}(\mathbf{a} + \mathbf{b})) \\
&= f_0(\mathbf{u}(\mathbf{a})) + \frac{\partial f_0}{\partial u_1} \left(\frac{\partial u_1}{\partial a_1} b_1 + \frac{\partial u_1}{\partial a_2} b_2 \right) \\
&\quad + \frac{\partial f_0}{\partial u_2} \left(\frac{\partial u_2}{\partial a_1} b_1 + \frac{\partial u_2}{\partial a_2} b_2 \right) + o(\|\mathbf{b}\|_{\mathbb{R}^2}) \\
&= f_0(\mathbf{u}(\mathbf{a})) + (p_1 \ p_2) \begin{pmatrix} \frac{\partial u_1}{\partial a_1} & \frac{\partial u_1}{\partial a_2} \\ \frac{\partial u_2}{\partial a_1} & \frac{\partial u_2}{\partial a_2} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + o(\|\mathbf{b}\|_{\mathbb{R}^2}). \tag{1.1.24}
\end{aligned}$$

On the other hand, if Eq. (1.1.11) is partially differentiated with respect to a_1 , then we have

$$\frac{\partial \mathbf{K}}{\partial a_1} \mathbf{u} + \mathbf{K} \frac{\partial \mathbf{u}}{\partial a_1} = \mathbf{0}_{\mathbb{R}^2},$$

which can be written as

$$\frac{e_Y}{l} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} + \frac{e_Y}{l} \begin{pmatrix} a_1 + a_2 & -a_2 \\ -a_2 & a_2 \end{pmatrix} \begin{pmatrix} \frac{\partial u_1}{\partial a_1} \\ \frac{\partial u_2}{\partial a_1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Hence, it follows that

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial a_1} &= -\mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial a_1} \mathbf{u} \\ &= - \begin{pmatrix} \frac{1}{a_1} & \frac{1}{a_1} \\ \frac{1}{a_1} & \frac{1}{a_1} + \frac{1}{a_2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} -\frac{u_1}{a_1} \\ -\frac{u_1}{a_1} \end{pmatrix}. \end{aligned} \quad (1.1.25)$$

Similarly, partially differentiating Eq. (1.1.11) with respect to a_2 yields

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial a_2} &= -\mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial a_2} \mathbf{u} \\ &= - \begin{pmatrix} \frac{1}{a_1} & \frac{1}{a_1} \\ \frac{1}{a_1} & \frac{1}{a_1} + \frac{1}{a_2} \end{pmatrix} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{u_2 - u_1}{a_2} \end{pmatrix}. \end{aligned} \quad (1.1.26)$$

Therefore, upon substituting Eqs. (1.1.25) and (1.1.26) into Eq. (1.1.24), we obtain

$$\begin{aligned} \tilde{f}_0(\mathbf{a} + \mathbf{b}) &= f_0(\mathbf{u}(\mathbf{a})) + (p_1 \ p_2) \begin{pmatrix} -\frac{u_1}{a_1} & 0 \\ -\frac{u_1}{a_1} & -\frac{u_2 - u_1}{a_2} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + o(\|\mathbf{b}\|_{\mathbb{R}^2}) \\ &= f_0(\mathbf{u}(\mathbf{a})) + \left(-\frac{u_1}{a_1} (p_1 + p_2) - \frac{u_2 - u_1}{a_2} p_2 \right) \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + o(\|\mathbf{b}\|_{\mathbb{R}^2}) \\ &= f_0(\mathbf{u}(\mathbf{a})) + \mathbf{g}_0 \cdot \mathbf{b} + o(\|\mathbf{b}\|_{\mathbb{R}^2}). \end{aligned} \quad (1.1.27)$$

If the solution of the state equation (Eq. (1.1.20)) is used in the previous equation, it becomes apparent that \mathbf{g}_0 of Eq. (1.1.27) agrees with Eq. (1.1.22). Moreover, if we use the notation $\varepsilon(u_1) = u_1/l$ and $\sigma(u_1) = e_Y \varepsilon(u_1)$ for the strain and stress, then we obtain

$$\mathbf{g}_0 = -\frac{e_Y}{l} \begin{pmatrix} u_1^2 \\ (u_2 - u_1)^2 \end{pmatrix} = l \begin{pmatrix} -\sigma(u_1) \varepsilon(u_1) \\ -\sigma(u_2 - u_1) \varepsilon(u_2 - u_1) \end{pmatrix}. \quad (1.1.28)$$

Equation (1.1.28) is an equation in which \mathbf{g}_0 is expressed as a function of the state variable \mathbf{u} .

Let us also find the second-order derivative of \tilde{f}_0 (the Hesse matrix of \tilde{f}_0) with respect to the variation of \mathbf{a} . If the chain rule of differentiation is used on \mathbf{g}_0 of Eq. (1.1.28), then we have

$$\begin{aligned}\mathbf{g}_0(\mathbf{a} + \mathbf{b}) &= \mathbf{g}_0(\mathbf{a}) + \frac{\partial \mathbf{g}_0}{\partial \mathbf{u}^\top} \frac{\partial \mathbf{u}}{\partial \mathbf{a}^\top} \mathbf{b} + o(\|\mathbf{b}\|_{\mathbb{R}^2}) \\ &= \mathbf{g}_0(\mathbf{a}) + \begin{pmatrix} \frac{\partial g_{01}}{\partial u_1} & \frac{\partial g_{01}}{\partial u_2} \\ \frac{\partial g_{02}}{\partial u_1} & \frac{\partial g_{02}}{\partial u_2} \end{pmatrix} \begin{pmatrix} \frac{\partial u_1}{\partial a_1} & \frac{\partial u_1}{\partial a_2} \\ \frac{\partial u_2}{\partial a_1} & \frac{\partial u_2}{\partial a_2} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + o(\|\mathbf{b}\|_{\mathbb{R}^2}).\end{aligned}$$

From this, the Hesse matrix of f_0 can be computed:

$$\begin{aligned}\mathbf{H}_0 &= \frac{\partial \mathbf{g}_0}{\partial \mathbf{u}^\top} \frac{\partial \mathbf{u}}{\partial \mathbf{a}^\top} \\ &= -\frac{e_Y}{l} \begin{pmatrix} 2u_1 & 0 \\ -2(u_2 - u_1) & 2(u_2 - u_1) \end{pmatrix} \begin{pmatrix} -\frac{u_1}{a_1} & 0 \\ -\frac{u_1}{a_1} - \frac{u_2 - u_1}{a_2} \end{pmatrix} \\ &= \frac{e_Y}{l} \begin{pmatrix} \frac{2u_1^2}{a_1} & 0 \\ 0 & \frac{2(u_2 - u_1)^2}{a_2} \end{pmatrix} \\ &= l \begin{pmatrix} \frac{2\sigma(u_1)\varepsilon(u_1)}{a_1} & 0 \\ 0 & \frac{2\sigma(u_2 - u_1)\varepsilon(u_2 - u_1)}{a_2} \end{pmatrix}. \quad (1.1.29)\end{aligned}$$

Using the solution of the state equation (Eq. (1.1.20)), it can easily be seen that the \mathbf{H}_0 of Eq. (1.1.29) agrees with that of Eq. (1.1.23). Here, although the \mathbf{H}_0 in Eq. (1.1.23) agrees with the partial derivative of \mathbf{g}_0 in Eq. (1.1.22) with respect to \mathbf{a}^\top , the \mathbf{H}_0 in Eq. (1.1.29) cannot be obtained from such a relationship. The reason for this is that the state variable is used in obtaining the \mathbf{H}_0 of Eq. (1.1.29).

1.1.6 The Adjoint Variable Method

Finally, let us find \mathbf{g}_0 through the adjoint variable method, which utilizes the Lagrange multiplier method. The details of the adjoint variable method are presented later on in Sect. 2.6.5 and so, for the moment, we shall limit ourselves to its formal application.

Let the Lagrange function for the cost function f_0 be

$$\begin{aligned}\mathcal{L}_0(\mathbf{a}, \mathbf{u}, \mathbf{v}_0) &= f_0(\mathbf{u}) + \mathcal{L}_S(\mathbf{a}, \mathbf{u}, \mathbf{v}_0) \\ &= \mathbf{p} \cdot \mathbf{u} - \mathbf{v}_0 \cdot (\mathbf{K}(\mathbf{a})\mathbf{u} - \mathbf{p}),\end{aligned}\quad (1.1.30)$$

where \mathcal{L}_S denotes a Lagrange function with respect to the state determination problem (Problem 1.1.3) defined by Eq. (1.1.12). Here, $\mathbf{v}_0 = (v_{01}, v_{02})^\top \in U = \mathbb{R}^2$ includes the subscript 0 in order to indicate that it is an adjoint variable (Lagrange multiplier) prepared for f_0 . Going forward, whenever f_i is a function of the state variable \mathbf{u} , the adjoint variable will be written as \mathbf{v}_i .

The adjoint variable method is a technique for finding \mathbf{g}_0 using the stationary conditions of $\mathcal{L}_0(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)$ with respect to arbitrary variations of \mathbf{u} and \mathbf{v}_0 . The (total) derivative of \mathcal{L}_0 with respect to an arbitrary variation $(\mathbf{b}, \hat{\mathbf{u}}, \hat{\mathbf{v}}_0) \in X \times U \times U$ of $(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)$ is

$$\begin{aligned}\mathcal{L}'_0(\mathbf{a}, \mathbf{u}, \mathbf{v}_0) [\mathbf{b}, \hat{\mathbf{u}}, \hat{\mathbf{v}}_0] &= \mathcal{L}_{0\mathbf{a}}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0) [\mathbf{b}] + \mathcal{L}_{0\mathbf{u}}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0) [\hat{\mathbf{u}}] + \mathcal{L}_{0\mathbf{v}_0}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0) [\hat{\mathbf{v}}_0].\end{aligned}\quad (1.1.31)$$

In this book, we use the notation $(\cdot)_a$ for $\partial(\cdot)/\partial a$. The third term on the right-hand side of Eq. (1.1.31) is

$$\mathcal{L}_{0\mathbf{v}_0}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0) [\hat{\mathbf{v}}_0] = \mathcal{L}_S(\mathbf{a}, \mathbf{u}, \hat{\mathbf{v}}_0). \quad (1.1.32)$$

Equation (1.1.32) is the Lagrange function with respect to the state determination problem (Problem 1.1.3) defined in Eq. (1.1.12) and, if \mathbf{u} is a solution of the state determination problem, the third term on the right-hand side of Eq. (1.1.31) is zero.

Moreover, the second term on the right-hand side of Eq. (1.1.31) is

$$\begin{aligned}\mathcal{L}_{0\mathbf{u}}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0) [\hat{\mathbf{u}}] &= f_{0\mathbf{u}}(\mathbf{u}) [\hat{\mathbf{u}}] + \mathcal{L}_{S\mathbf{u}}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0) [\hat{\mathbf{u}}] \\ &= \mathbf{p} \cdot \hat{\mathbf{u}} - \mathbf{v}_0 \cdot (\mathbf{K}(\mathbf{a})\hat{\mathbf{u}}) \\ &= \hat{\mathbf{u}} \cdot (\mathbf{p} - \mathbf{K}^\top(\mathbf{a})\mathbf{v}_0).\end{aligned}\quad (1.1.33)$$

Here, if \mathbf{v}_0 can be determined so that Eq. (1.1.33) is zero for arbitrary $\hat{\mathbf{u}} \in U$, then the second term on the right-hand side of Eq. (1.1.31) also vanishes. The condition here is equivalent to setting \mathbf{v}_0 to be the solution of the following adjoint problem.

Problem 1.1.5 (Adjoint Problem with Respect to f_0) Let $\mathbf{K}(\mathbf{a})$ and \mathbf{p} be as in Problem 1.1.3 and find $\mathbf{v}_0 \in U$ satisfying

$$\mathbf{K}^\top(\mathbf{a})\mathbf{v}_0 = \mathbf{p}. \quad (1.1.34)$$

□

Upon comparison of Problem 1.1.3 and Problem 1.1.5, using the fact that $\mathbf{K}^\top = \mathbf{K}$, we obtain

$$\mathbf{v}_0 = \mathbf{u}. \quad (1.1.35)$$

As in the above equation, the relationship where the state variable is equal to the adjoint variable is called a self-adjoint relationship. In fact, the right-hand side of Eq. (1.1.34) is $\partial f_0(\mathbf{u}) / \partial \mathbf{u}$. In Problem 1.1.4, $f_0(\mathbf{u}) = \mathbf{p} \cdot \mathbf{u}$, and the self-adjoint relationship holds. That is, the self-adjoint property holds when f_0 is selected so that $\partial f_0(\mathbf{u}) / \partial \mathbf{u}$ is equal to the right-hand side of the state equation (Eq. (1.1.11)). In generality, state and adjoint equations are different and, in such a case, their relationship is said to be non-self adjoint. An example of this is presented in Practice 1.1. We would now like to consider the meaning of the adjoint equation, and we remark that the following arguments apply to other adjoint equations.

The first term on the right-hand side of Eq. (1.1.31) can be obtained:

$$\begin{aligned} \mathcal{L}_{0a}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\mathbf{b}] &= - \left\{ \mathbf{v}_0 \cdot \left(\frac{\partial \mathbf{K}(\mathbf{a})}{\partial a_1} \mathbf{u} \frac{\partial \mathbf{K}(\mathbf{a})}{\partial a_2} \mathbf{u} \right) \right\} \mathbf{b} \\ &= -\frac{eY}{l} \left\{ (v_{01} \ v_{02}) \left(\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \right) \right\} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \\ &= -\frac{eY}{l} (v_{01} \ v_{02}) \begin{pmatrix} u_1 & u_1 - u_2 \\ 0 & u_2 - u_1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \\ &= -\frac{eY}{l} (u_1 v_{01} (u_2 - u_1) (v_{02} - v_{01})) \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \\ &= l (-\sigma(u_1) \varepsilon(v_{01}) - \sigma(u_2 - u_1) \varepsilon(v_{02} - v_{01})) \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \\ &= \mathbf{g}_0 \cdot \mathbf{b}. \end{aligned} \quad (1.1.36)$$

Here we remark that \mathbf{g}_0 matches the results from the direct differentiation method (Eq. (1.1.28)).

Based on the above results, if \mathbf{u} and \mathbf{v}_0 are solutions of Problem 1.1.3 and Problem 1.1.5, respectively, the second and third terms on the right-hand side of Eq. (1.1.31) are zero, and the following equation holds:

$$\mathcal{L}'_0(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\mathbf{b}, \hat{\mathbf{u}}, \hat{\mathbf{v}}_0] = \mathcal{L}_{0a}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\mathbf{b}] = \tilde{f}'_0(\mathbf{a})[\mathbf{b}] = \mathbf{g}_0 \cdot \mathbf{b}. \quad (1.1.37)$$

The method of obtaining the derivative of the cost function when equality constraints (state equations) are satisfied in this way is shown in Sect. 2.6, where

Eq. (1.1.36) corresponds to Eq. (2.6.25). It should, however, be noted that in Sect. 2.6, the gradient of \tilde{f} is written as \tilde{g} .

Let us also examine the relationship between the Lagrange function \mathcal{L}_0 and the Hesse matrix \mathbf{H}_0 .

As explained in Sect. 2.1, if an optimal design problem is to be replaced by an optimization problem in which design variables and state variables are not distinguished, and are treated as variables, then the variable of the optimization problem becomes a combination of the state and design variables of the optimal design problem. Following this approach here, the design variable of the optimization problem is set to be $\mathbf{x} = (\mathbf{a}^\top, \mathbf{u}^\top)^\top \in \mathbb{R}^4$. In order to simplify the notation, $(\mathbf{a}^\top, \mathbf{u}^\top)^\top$ will be written as (\mathbf{a}, \mathbf{u}) . The Lagrange multiplier with respect to the equality constraint will be written as \mathbf{v}_0 , and the second-order derivative of the Lagrange function \mathcal{L}_0 with respect to arbitrary variations $(\mathbf{b}_2, \hat{\mathbf{u}}_2) \in X \times U$ and $(\mathbf{b}_1, \hat{\mathbf{u}}_1) \in X \times U$ of the design variables (\mathbf{a}, \mathbf{u}) will be written $\mathcal{L}_{0(\mathbf{a}, \mathbf{u}), (\mathbf{a}, \mathbf{u})}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[(\mathbf{b}_1, \hat{\mathbf{u}}_1), (\mathbf{b}_2, \hat{\mathbf{u}}_2)]$. In this case, we have

$$\begin{aligned}
& \mathcal{L}_{0(\mathbf{a}, \mathbf{u}), (\mathbf{a}, \mathbf{u})}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[(\mathbf{b}_1, \hat{\mathbf{u}}_1), (\mathbf{b}_2, \hat{\mathbf{u}}_2)] \\
&= (\mathcal{L}_{0\mathbf{a}}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\mathbf{b}_1] + \mathcal{L}_{0\mathbf{u}}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\hat{\mathbf{u}}_1])_{\mathbf{a}}[\mathbf{b}_2] \\
&\quad + (\mathcal{L}_{0\mathbf{a}}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\mathbf{b}_1] + \mathcal{L}_{0\mathbf{u}}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\hat{\mathbf{u}}_1])_{\mathbf{u}}[\hat{\mathbf{u}}_2] \\
&= (f_{0\mathbf{u}} \cdot \hat{\mathbf{u}}_1 + \mathcal{L}_{\mathbf{S}\mathbf{a}}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\mathbf{b}_1] + \mathcal{L}_{\mathbf{S}\mathbf{u}}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\hat{\mathbf{u}}_1])_{\mathbf{a}}[\mathbf{b}_2] \\
&\quad + (f_{0\mathbf{u}} \cdot \hat{\mathbf{u}}_1 + \mathcal{L}_{\mathbf{S}\mathbf{a}}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\mathbf{b}_1] + \mathcal{L}_{\mathbf{S}\mathbf{u}}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\hat{\mathbf{u}}_1])_{\mathbf{u}}[\hat{\mathbf{u}}_2] \\
&= \begin{pmatrix} \mathbf{b}_2 \\ \hat{\mathbf{u}}_2 \end{pmatrix} \cdot \left(\mathbf{H}_{\mathcal{L}_{\mathbf{S}}} \begin{pmatrix} \mathbf{b}_1 \\ \hat{\mathbf{u}}_1 \end{pmatrix} \right), \tag{1.1.38}
\end{aligned}$$

where

$$\mathbf{H}_{\mathcal{L}_{\mathbf{S}}} = \begin{pmatrix} \mathcal{L}_{\mathbf{S}\mathbf{aa}} & \mathcal{L}_{\mathbf{S}\mathbf{au}} \\ \mathcal{L}_{\mathbf{S}\mathbf{ua}} & \mathcal{L}_{\mathbf{S}\mathbf{uu}} \end{pmatrix} = - \begin{pmatrix} \mathbf{0}_{\mathbb{R}^{2 \times 2}} & \begin{pmatrix} \mathbf{v}_0^\top \mathbf{K}_{a_1} \\ \mathbf{v}_0^\top \mathbf{K}_{a_2} \end{pmatrix} \\ \begin{pmatrix} \mathbf{K}_{a_1}^\top \mathbf{v}_0 & \mathbf{K}_{a_2}^\top \mathbf{v}_0 \end{pmatrix} & \mathbf{0}_{\mathbb{R}^{2 \times 2}} \end{pmatrix}. \tag{1.1.39}$$

From Eq. (1.1.39), it is apparent that the matrix $\mathbf{H}_{\mathcal{L}_{\mathbf{S}}}$ need not be positive definite.

Here \mathbf{u} and \mathbf{v}_0 denote, respectively, the solutions to the state determination problem (Problem 1.1.3) and the adjoint problem (Problem 1.1.5), subject to a design variable \mathbf{a} . Furthermore, we assume that $\hat{\mathbf{v}}$ (the letter \mathbf{v} is a bold Greek upsilon) denotes a variation of \mathbf{u} under the equality constraint of the state determination problem corresponding to an arbitrary variation $\mathbf{b} \in X$ of \mathbf{a} . In Chap. 2, we call the set of $(\mathbf{b}, \hat{\mathbf{v}})$ the feasible direction set or the tangent plane on $X \times U$ satisfying the equality constraint of the state determination problem (see $T_V(\mathbf{x})$ in Eq. (2.6.2), Theorems 2.6.6 and 2.6.7). Here, the cross-sectional derivative of the Lagrange

function with respect to the state determination problem is

$$\mathcal{L}_{S(a,u)}(\mathbf{a}, \mathbf{u}, \mathbf{v}) [\mathbf{b}, \hat{\mathbf{v}}] = \mathbf{v} \cdot \{ -(\mathbf{K}'(\mathbf{a})[\mathbf{b}])\mathbf{u} - \mathbf{K}(\mathbf{a})\hat{\mathbf{v}} \} = 0. \quad (1.1.40)$$

This yields the identity

$$\hat{\mathbf{v}} = -\mathbf{K}^{-1}(\mathbf{a}) (\mathbf{K}'(\mathbf{a})[\mathbf{b}]) = \begin{pmatrix} -\frac{u_1}{a_1} & 0 \\ -\frac{u_1}{a_1} - \frac{u_2 - u_1}{a_2} & \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}. \quad (1.1.41)$$

Equation (1.1.41) is equivalent to the conditions expressed by Eqs. (1.1.25) and (1.1.26). Moreover, using the self-adjoint relationship, Eq. (1.1.38) becomes

$$\begin{aligned} & \mathcal{L}_{0(a,u),(a,u)}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0) [(\mathbf{b}_1, \hat{\mathbf{v}}_1), (\mathbf{b}_2, \hat{\mathbf{v}}_2)] \\ &= \mathcal{L}_{Sau}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0) [\mathbf{b}_1, \hat{\mathbf{v}}_2] + \mathcal{L}_{Sua}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0) [\mathbf{b}_2, \hat{\mathbf{v}}_1] \\ &= -\begin{pmatrix} b_{11} \\ b_{12} \end{pmatrix} \cdot \left\{ \begin{pmatrix} \mathbf{v}_0^\top \mathbf{K}_{a_1} \\ \mathbf{v}_0^\top \mathbf{K}_{a_2} \end{pmatrix} \begin{pmatrix} -u_1/a_1 & 0 \\ -u_1/a_1 - (u_2 - u_1)/a_2 \end{pmatrix} \right. \\ & \quad \left. + \begin{pmatrix} \mathbf{v}_0^\top \mathbf{K}_{a_1} \\ \mathbf{v}_0^\top \mathbf{K}_{a_2} \end{pmatrix} \begin{pmatrix} -u_1/a_1 & 0 \\ -u_1/a_1 - (u_2 - u_1)/a_2 \end{pmatrix}^\top \right\} \begin{pmatrix} b_{21} \\ b_{22} \end{pmatrix} \\ &= \mathbf{b}_1 \cdot (\mathbf{H}_0 \mathbf{b}_2). \end{aligned}$$

This shows that the second-order cross-sectional derivative of \tilde{f}_0 agrees with Eq. (1.1.29) and is expressed as

$$h_0(\mathbf{a}) [\mathbf{b}_1, \mathbf{b}_2] = \mathbf{b}_1 \cdot (\mathbf{H}_0 \mathbf{b}_2). \quad (1.1.42)$$

The above results clarify that the Hesse matrix of the Lagrange function \mathcal{L}_0 with respect to an arbitrary variation of (\mathbf{a}, \mathbf{u}) agrees with the Hesse matrix $\mathbf{H}_{\mathcal{L}_S}$ of \mathcal{L}_S , and that it is not necessarily positive definite. However, when we assume that \mathbf{u} denotes the solution of the state determination problem as the design variable is varied and that $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{u}}_2$ denote variations of \mathbf{u} , we showed that the Hesse matrix with respect to arbitrary variations $\mathbf{b}_1, \mathbf{b}_2 \in X$ of \mathbf{a} from \mathcal{L}_0 is the same as the Hesse matrix \mathbf{H}_0 obtained via the other methods.

In the optimal design problem (Problem 1.1.4) considered in this section, we have assumed that the design variable $\mathbf{a} \in \mathbb{R}^2$ is a cross-sectional area. Therefore, $\mathbf{K}(\mathbf{a})$ was a linear form of \mathbf{a} in Eq. (1.1.11) and hence, $\mathcal{L}_{Saa} = \mathbf{0}_{\mathbb{R}^{2 \times 2}}$ in Eqs. (1.1.39) and (1.1.42). However, if the design variable $\mathbf{a} \in \mathbb{R}^2$ is assumed to be the length of one side of a square cross-section, then $\mathbf{K}(\mathbf{a})$ becomes $\mathbf{K}(\mathbf{a}^2)$ and $\mathcal{L}_{Saa} \neq \mathbf{0}_{\mathbb{R}^{2 \times 2}}$ (see Practice 1.5). In this way, we see that \mathcal{L}_{0aa} may not be $\mathbf{0}_{\mathbb{R}^{2 \times 2}}$, depending on the choice of design variables or cost functions. Nevertheless, when the state determination problem is linear the condition $\mathcal{L}_{Suu} = \mathbf{0}_{\mathbb{R}^{2 \times 2}}$ always holds.

In the method used above, Eq. (1.1.41) was utilized to obtain the second cross-sectional derivative from the first cross-sectional derivative. Equation (1.1.41) accords with Eqs. (1.1.25) and (1.1.26) from which we see that the direct differentiation method was actually applied. These results do not always hold in general, such as in those problems given in Chaps. 8 and 9. In these cases, the Lagrange multiplier method can be used to obtain the second-order derivative of cost functions. Such a method will also be described in the succeeding discussions. However, in cases where the second-order derivatives are used in solving the optimization problems, a key idea, whose detail will be presented in Chap. 3, is required.

In Chap. 4, the Fréchet derivative will be defined as a generalized derivative (Definition 4.5.4). Following the definition of the second-order derivative, here, we fix \mathbf{b}_1 and consider differentiating the first cross-sectional derivative $\tilde{f}'_0(\mathbf{a})[\mathbf{b}_1] = \mathbf{g}_0 \cdot \mathbf{b}_1$. To do this, we define the Lagrange function for $\mathbf{g}_0 \cdot \mathbf{b}_1$ by

$$\mathcal{L}_{10}(\mathbf{a}, \mathbf{u}, \mathbf{w}_0) = \mathbf{g}_0(\mathbf{u}) \cdot \mathbf{b}_1 + \mathcal{L}_S(\mathbf{a}, \mathbf{u}, \mathbf{w}_0), \quad (1.1.43)$$

where $\mathbf{g}_0(\mathbf{u})$ and \mathcal{L}_S are given by Eqs. (1.1.12) and (1.1.36), respectively. $\mathbf{w}_0 = (w_{01}, w_{02})^\top$ is the adjoint variable provided for \mathbf{u} in $\mathbf{g}_0(\mathbf{u})$ satisfying the state determination problem.

With respect to arbitrary variations $(\mathbf{b}_2, \hat{\mathbf{u}}, \hat{\mathbf{w}}_0) \in X \times U^2$ of $(\mathbf{a}, \mathbf{u}, \mathbf{w}_0)$, the derivative of \mathcal{L}_{10} is written as

$$\begin{aligned} \mathcal{L}'_{10}(\mathbf{a}, \mathbf{u}, \mathbf{w}_0) & [\mathbf{b}_2, \hat{\mathbf{u}}, \hat{\mathbf{w}}_0] \\ &= \mathcal{L}_{10a}(\mathbf{a}, \mathbf{u}, \mathbf{w}_0)[\mathbf{b}_2] + \mathcal{L}_{10u}(\mathbf{a}, \mathbf{u}, \mathbf{w}_0)[\hat{\mathbf{u}}] \\ &\quad + \mathcal{L}_{10w_0}(\mathbf{a}, \mathbf{u}, \mathbf{w}_0)[\hat{\mathbf{w}}_0]. \end{aligned} \quad (1.1.44)$$

The third term on the right-hand side of Eq. (1.1.44) vanishes if \mathbf{u} is the solution of the state determination problem. Moreover, the second term on the right-hand side of Eq. (1.1.44) is

$$\begin{aligned} \mathcal{L}_{10u}(\mathbf{a}, \mathbf{u}, \mathbf{w}_0)[\hat{\mathbf{u}}] &= \mathbf{g}_{0u^\top}(\mathbf{u})[\hat{\mathbf{u}}] \cdot \mathbf{b}_1 + \mathcal{L}_{Su}(\mathbf{a}, \mathbf{u}, \mathbf{w}_0)[\hat{\mathbf{u}}] \\ &= \hat{\mathbf{u}} \cdot \mathbf{q} - \mathbf{w}_0 \cdot (\mathbf{K}(\mathbf{a}) \hat{\mathbf{u}}) \\ &= \hat{\mathbf{u}} \cdot (\mathbf{q} - \mathbf{K}^\top(\mathbf{a}) \mathbf{w}_0), \end{aligned} \quad (1.1.45)$$

where

$$\mathbf{q} = \mathbf{g}_{0u}^\top(\mathbf{u}) \mathbf{b}_1 = -\frac{2e_Y}{l} \begin{pmatrix} u_1 & u_1 - u_2 \\ 0 & u_2 - u_1 \end{pmatrix} \begin{pmatrix} b_{11} \\ b_{12} \end{pmatrix}. \quad (1.1.46)$$

Here, the condition that Eq. (1.1.45) is zero for arbitrary $\hat{\mathbf{u}} \in U$ is equivalent to setting \mathbf{w}_0 to be the solution of the following adjoint problem.

Problem 1.1.6 (Adjoint Problem with Respect to $\mathbf{g}_0(\mathbf{u}) \cdot \mathbf{b}_1$) Let $\mathbf{K}(\mathbf{a})$ be as in Problem 1.1.3, and \mathbf{q} be given by Eq. (1.1.46). Find $\mathbf{w}_0 \in U$ satisfying

$$\mathbf{K}^\top(\mathbf{a}) \mathbf{w}_0 = \mathbf{q}. \quad \square$$

The solution of Problem 1.1.6 is

$$\mathbf{w}_0 = \left(\mathbf{K}^\top(\mathbf{a}) \right)^{-1} \mathbf{g}_{0\mathbf{u}}^\top(\mathbf{u}) \mathbf{b}_1 = -2 \begin{pmatrix} \frac{u_1}{a_1} & 0 \\ \frac{u_1}{a_1} & \frac{u_2 - u_1}{a_2} \end{pmatrix} \begin{pmatrix} b_{11} \\ b_{12} \end{pmatrix}. \quad (1.1.47)$$

Here, \mathbf{w}_0 is a function of \mathbf{b}_1 , and so is written as $\mathbf{w}_0(\mathbf{b}_1)$.

Finally, the first term on the right-hand side of Eq. (1.1.44) becomes

$$\begin{aligned} \mathcal{L}_{10\mathbf{a}}(\mathbf{a}, \mathbf{u}, \mathbf{w}_0(\mathbf{b}_1))[\mathbf{b}_2] \\ = - \left\{ \mathbf{w}_0(\mathbf{b}_1) \cdot \left(\frac{\partial \mathbf{K}(\mathbf{a})}{\partial a_1} \mathbf{u} \frac{\partial \mathbf{K}(\mathbf{a})}{\partial a_2} \mathbf{u} \right) \right\} \mathbf{b}_2. \end{aligned} \quad (1.1.48)$$

Substituting Eq. (1.1.47) into Eq. (1.1.48), we obtain

$$\begin{aligned} \mathcal{L}_{10\mathbf{a}}(\mathbf{a}, \mathbf{u}_0, \mathbf{w}_0(\mathbf{b}_1))[\mathbf{b}_2] \\ = h_0(\mathbf{a})[\mathbf{b}_1, \mathbf{b}_2] = \mathbf{b}_1 \cdot (\mathbf{H}_0 \mathbf{b}_2) = \mathbf{g}_{H0}(\mathbf{a}, \mathbf{b}_1) \cdot \mathbf{b}_2, \end{aligned} \quad (1.1.49)$$

where

$$\mathbf{g}_{H0}(\mathbf{a}, \mathbf{b}_1) = \mathcal{L}_{10\mathbf{a}}(\mathbf{a}, \mathbf{u}, \mathbf{w}_0(\mathbf{b}_1)). \quad (1.1.50)$$

In this book, \mathbf{g}_{H0} is called the Hesse gradient.

1.1.7 Optimality Conditions

The previous section explored methods for calculating \mathbf{g}_0 (the gradient of \tilde{f}_0 with respect to the variation of the cross-section $\mathbf{a} \in \mathcal{D}^\circ$), the Hesse matrix \mathbf{H}_0 , and the gradient \mathbf{g}_1 of f_1 . We now return to Problem 1.1.4 and consider optimality conditions that the optimal cross-section satisfies.

As described in Sect. 1.1.4 the convexity of \tilde{f}_0 was obtained from the fact that the Hesse matrix \mathbf{H}_0 is positive definite on X (Theorem 2.4.6). We also showed that f_1 is a linear function of \mathbf{a} and that it is therefore convex on X (Theorem 2.4.4). Problem 1.1.4 is then a convex optimization problem and, as will be shown later, a design variable $\mathbf{a} \in \mathcal{D}^\circ$ which satisfies the Karush–Kuhn–Tucker conditions

(Theorem 2.7.5) is the minimizer of Problem 1.1.4 (Theorem 2.7.9). Let us now find the KKT conditions for Problem 1.1.4.

Let the Lagrange function for Problem 1.1.4 be

$$\mathcal{L}(\mathbf{a}, \lambda_1) = \tilde{f}_0(\mathbf{a}) + \lambda_1 f_1(\mathbf{a}),$$

where $\lambda_1 \in \mathbb{R}$ is a Lagrange multiplier with respect to $f_1(\mathbf{a}) \leq 0$. Then the KKT conditions for Problem 1.1.4 are given by

$$\mathcal{L}_{\mathbf{a}}(\mathbf{a}, \lambda_1) = \mathbf{g}_0 + \lambda_1 \mathbf{g}_1 = \mathbf{0}_{\mathbb{R}^2}, \quad (1.1.51)$$

$$\mathcal{L}_{\lambda_1}(\mathbf{a}, \lambda_1) = f_1(\mathbf{a}) = l(a_1 + a_2) - c_1 \leq 0, \quad (1.1.52)$$

$$\lambda_1 f_1(\mathbf{a}) = 0, \quad (1.1.53)$$

$$\lambda_1 \geq 0. \quad (1.1.54)$$

A detailed explanation regarding the meaning of the KKT conditions will be deferred until Sect. 2.7.3, but let us now take a look into their general meaning.

First of all, when the cross-section is optimal, Eqs. (1.1.51) and (1.1.54) describe a trade-off relationship between the objective and the constraint functions. In fact, upon taking the inner product of \mathbf{b} with both sides of Eq. (1.1.51), we can obtain

$$\lambda_1 = -\frac{\mathbf{g}_0 \cdot \mathbf{b}}{\mathbf{g}_1 \cdot \mathbf{b}}. \quad (1.1.55)$$

The numerator and the denominator on the right-hand side of Eq. (1.1.55) represent the amount of variation in f_1 and f_0 when the design variable is varied by \mathbf{b} . Here, $\lambda_1 > 0$ indicates the fact that the signs of the variations differ. In other words, there is a trade-off relationship between f_1 and f_0 .

Finally, we remark that Eq. (1.1.52) is the original constraint condition. Also, we say that Eq. (1.1.53) is a complementarity condition. If an inequality constraint can be satisfied by an equality (referred to as active), then Eq. (1.1.53) allows $\lambda_1 > 0$. Similarly, if it can be satisfied as an inequality (referred to as inactive) then $\lambda_1 = 0$ and this acts to inactivate the constraint.

Next let us consider the physical interpretation of Eq. (1.1.51). If \mathbf{g}_0 from Eq. (1.1.36) and \mathbf{g}_1 from Eq. (1.1.17) are substituted into Eq. (1.1.51), then we obtain

$$l \begin{pmatrix} -\sigma(u_1) \varepsilon(u_1) \\ -\sigma(u_2 - u_1) \varepsilon(u_2 - u_1) \end{pmatrix} + \lambda_1 l \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

This equation implies that

$$\sigma(u_1) \varepsilon(u_1) = \sigma(u_2 - u_1) \varepsilon(u_2 - u_1) = \lambda_1. \quad (1.1.56)$$

In other words, when Problem 1.1.4 is minimized, the strain energy densities (w of Eq. (1.1.7)) of the two elastic bodies agree and λ_1 is twice the strain energy density. Therefore, λ_1 is greater than zero when \mathbf{p} generates a non-zero stress on the two one-dimensional elastic bodies, and the volume constraint is active at the minimizer.

1.1.8 Numerical Example

Let us consider a concrete example and try to obtain the minimizer.

Exercise 1.1.7 (Mean Compliance Minimization) Find the minimizer \mathbf{a} in Problem 1.1.4, subject to $l = 1$, $e_Y = 1$, $c_1 = 1$, $\mathbf{p} = (1, 1)^\top$ and $\mathbf{a}_0 = (0.1, 0.1)^\top$. \square

Answer Substituting $l = 1$, $e_Y = 1$ and $\mathbf{p} = (1, 1)^\top$ into Eq. (1.1.21) gives

$$\tilde{f}_0(\mathbf{a}) = \frac{4}{a_1} + \frac{1}{a_2}. \quad (1.1.57)$$

Figure 1.4 shows \tilde{f}_0 . From Eqs. (1.1.22) and (1.1.17), the cross-sectional-area derivative of \tilde{f}_0 and f_1 are given by

$$\mathbf{g}_0 = -\begin{pmatrix} 4/a_1^2 \\ 1/a_2^2 \end{pmatrix}, \quad \mathbf{g}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

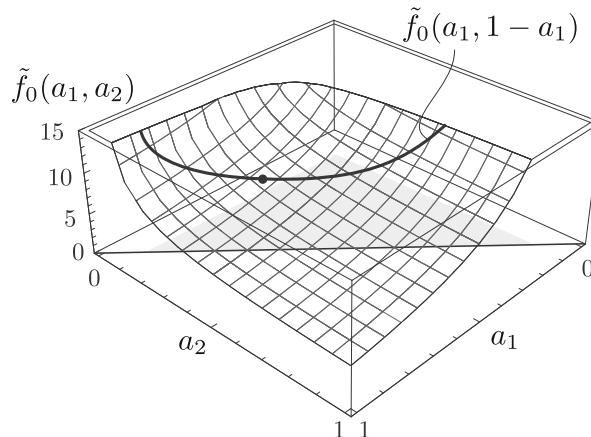


Fig. 1.4 Numerical example of the mean compliance minimization problem

If \mathbf{a} is a minimizer, then from Eq. (1.1.56) we have

$$\lambda_1 = \frac{4}{a_1^2} = \frac{1}{a_2^2}.$$

If \mathbf{a} is an element of \mathcal{D}° (see Eq. (1.1.16)), then λ_1 is positive and the complementarity condition allows for the inequality constraint with respect to f_1 to be satisfied with an equality. Here, if $a_2 = 1 - a_1$ is substituted into Eq. (1.1.57), then a_1 can be obtained from the stationary condition of \tilde{f}_0 with respect to an arbitrary variation of a_1 :

$$\frac{d}{da_1} \tilde{f}_0(a_1, 1 - a_1) = \frac{1}{(1 - a_1)^2} - \frac{4}{a_1^2} = 0.$$

We then obtain a_2 from $1 - a_1$:

$$\mathbf{a} = \begin{pmatrix} 2/3 \\ 1/3 \end{pmatrix}, \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

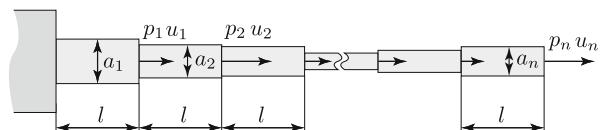
Among these values, $\mathbf{a} = (2/3, 1/3)^\top$ satisfies $\mathbf{a} \geq \mathbf{a}_0$. Due to the fact that \tilde{f}_0 and f_1 are convex functions, Problem 1.1.4 becomes a convex optimization problem, and from Theorem 2.7.9, the value of \mathbf{a} which satisfies the KKT condition is the minimizer of the problem. \square

1.2 Comparison of the Direct Differentiation Method and the Adjoint Variable Method

In Sect. 1.1, we considered how to find optimality conditions for Problem 1.1.4. The substitution method, the direct differentiation method, and the adjoint variable method were used to find the cross-sectional derivative of the cost function. Since we will later deal with optimization problems where the design variable is a function, and because the substitution method becomes quite complex in such a setting, we will exclude this method and compare the characteristics and applicable range of the direct differentiation and adjoint variable methods.

Let us consider a one-dimensional linear elastic body such as the one shown in Fig. 1.5. Here, the number of cross-sections in Problem 1.1.4 has been extended to

Fig. 1.5 A one-dimensional linear elastic body with n cross-sections



$n \in \mathbb{N}$ (the set of all natural numbers). The linear elasticity problem in this case is as follows.

Problem 1.2.1 (Multi-stepped 1D Linear Elastic Body) When $l \in \mathbb{R}$, $e_Y \in \mathbb{R}$, $\mathbf{a} \in \mathbb{R}^n$ ($\mathbf{a} \geq \mathbf{a}_0 > \mathbf{0}_{\mathbb{R}^n}$) and $\mathbf{p} \in \mathbb{R}^n$ are given with respect to the one-dimensional linear elastic body in Fig. 1.5, find $\mathbf{u} \in \mathbb{R}^n$ satisfying

$$\mathbf{K}(\mathbf{a})\mathbf{u} = \mathbf{p}. \quad (1.2.1)$$

Here, $\mathbf{K}(\mathbf{a})$ is an extension matrix of $\mathbf{K}(\mathbf{a})$ from Problem 1.1.3 (see Practice 1.5). \square

With respect to $\mathbf{u} \in \mathbb{R}^n$ and the Lagrange multiplier $\mathbf{v} \in \mathbb{R}^n$, the Lagrange function with respect to the state determination problem (Problem 1.2.1) is

$$\mathcal{L}_S(\mathbf{a}, \mathbf{u}, \mathbf{v}) = \mathbf{v} \cdot (-\mathbf{K}(\mathbf{a})\mathbf{u} + \mathbf{p}). \quad (1.2.2)$$

Here, Problem 1.2.1 is equivalent to finding $\mathbf{u} \in \mathbb{R}^n$ satisfying

$$\mathcal{L}_S(\mathbf{a}, \mathbf{u}, \mathbf{v}) = 0,$$

for all $\mathbf{v} \in \mathbb{R}^n$.

The number of constraint functions is set to be $m \in \mathbb{N}$, and the optimal design problem is as follows. In the following problem, we let $X = \mathbb{R}^n$, $U = \mathbb{R}^n$, and for a constant vector $\mathbf{a}_0 > \mathbf{0}_{\mathbb{R}^n}$ we set

$$\mathcal{D} = \{\mathbf{a} \in X \mid \mathbf{a} \geq \mathbf{a}_0\}. \quad (1.2.3)$$

Problem 1.2.2 (Multi-design Variable Multi-constraint) Let $X = U = \mathbb{R}^n$, and \mathcal{D} be given by Eq. (1.2.3). Also assume that a function $f_i : X \times U \rightarrow \mathbb{R}$ is given for each $i \in \{0, 1, \dots, m\}$. Given these conditions, find \mathbf{a} satisfying

$$\min_{(\mathbf{a}, \mathbf{u}) \in \mathcal{D} \times U} \{f_0(\mathbf{a}, \mathbf{u}) \mid f_1(\mathbf{a}, \mathbf{u}) \leq 0, \dots, f_m(\mathbf{a}, \mathbf{u}) \leq 0, \text{ Problem 1.2.1}\}. \quad \square$$

Let us use this problem to compare the method of direct differentiation to the adjoint variable approach, while formalizing techniques for calculating the cross-sectional-area gradients $\mathbf{g}_0, \dots, \mathbf{g}_m$ of the cost functions f_0, \dots, f_m . Hereinafter, $i \in \{0, 1, \dots, m\}$ is the subscript of the cost function f_i , and $j \in \{1, \dots, n\}$ is the subscript of the design variable a_j .

1.2.1 The Direct Differentiation Method

Let us first look at the method for calculating \mathbf{g}_i using the direct differentiation method. The solution of Problem 1.2.1 corresponding to $\mathbf{a} + \mathbf{b}$ with respect to an arbitrary $\mathbf{b} \in \mathbb{R}^n$ will be written as $\mathbf{u}(\mathbf{a} + \mathbf{b})$.

From the Taylor expansion of $\tilde{f}_i(\mathbf{a})$ about \mathbf{a} and the chain rule of differentiation, one can write

$$\begin{aligned}
& \tilde{f}_i(\mathbf{a} + \mathbf{b}) \\
&= f_i(\mathbf{a}, \mathbf{u}(\mathbf{a})) + \left(\frac{\partial f_i}{\partial a_1} \frac{\partial f_i}{\partial a_2} \cdots \frac{\partial f_i}{\partial a_n} \right) \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \\
&+ \left(\frac{\partial f_i}{\partial u_1} \frac{\partial f_i}{\partial u_2} \cdots \frac{\partial f_i}{\partial u_n} \right) \begin{pmatrix} \frac{\partial u_1}{\partial a_1} \frac{\partial u_1}{\partial a_2} \cdots \frac{\partial u_1}{\partial a_n} \\ \frac{\partial u_2}{\partial a_1} \frac{\partial u_2}{\partial a_2} \cdots \frac{\partial u_2}{\partial a_n} \\ \vdots \\ \frac{\partial u_n}{\partial a_1} \frac{\partial u_n}{\partial a_2} \cdots \frac{\partial u_n}{\partial a_n} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \\
&+ o(\|\mathbf{b}\|_{\mathbb{R}^n}) \\
&= f_i(\mathbf{a}, \mathbf{u}(\mathbf{a})) + f_{ia} \cdot \mathbf{b} + f_{iu} \cdot (\mathbf{u}_{a^\top} \mathbf{b}) + o(\|\mathbf{b}\|_{\mathbb{R}^n}) \\
&= f_i(\mathbf{a}, \mathbf{u}(\mathbf{a})) + \left\{ f_{ia} + (\mathbf{u}_{a^\top})^\top f_{iu} \right\} \cdot \mathbf{b} + o(\|\mathbf{b}\|_{\mathbb{R}^n}) \\
&= f_i(\mathbf{a}, \mathbf{u}(\mathbf{a})) + \mathbf{g}_i \cdot \mathbf{b} + o(\|\mathbf{b}\|_{\mathbb{R}^n}). \tag{1.2.4}
\end{aligned}$$

In this book, the matrix $(\partial u_i / \partial a_j)_{ij}$, consisting of the partial derivative of the column vector \mathbf{u} with respect to a row vector \mathbf{a}^\top will be written as \mathbf{u}_{a^\top} . If $f_i(\mathbf{a}, \mathbf{u})$ is given in Eq. (1.2.4), then f_{ia} and f_{iu} are known, so let us now consider a method for calculating \mathbf{u}_{a^\top} in order to find \mathbf{g}_i .

Taking the partial derivative of the state equation with respect to a_j yields

$$\frac{\partial \mathbf{K}}{\partial a_j} \mathbf{u} + \mathbf{K}(\mathbf{a}) \frac{\partial \mathbf{u}}{\partial a_j} = \mathbf{0}_{\mathbb{R}^n},$$

where $j \in \{1, \dots, n\}$, or equivalently

$$\frac{\partial \mathbf{u}}{\partial a_j} = -\mathbf{K}^{-1}(\mathbf{a}) \frac{\partial \mathbf{K}}{\partial a_j} \mathbf{u}. \quad (1.2.5)$$

Arranging Eq. (1.2.5) in rows with respect to $j \in \{1, \dots, n\}$ establishes $\mathbf{u}_{\mathbf{a}^\top}$.

The following statements can be made from the above observations.

Remark 1.2.3 (Characteristics of the Direct Differentiation Method) Compared with the adjoint variable method, the direct differentiation method has the following properties:

- (1) The direct differentiation method is effective when the number of cost functions is large, i.e., $m \gg 1$. This is because, once $\mathbf{u}_{\mathbf{a}^\top}$ has been computed from Eq. (1.2.5), $\mathbf{u}_{\mathbf{a}^\top}$ can be used for each of the cost functions f_0, \dots, f_m .
- (2) When the number of design variables is large ($n \gg 1$), the direct differentiation method becomes ineffective. This is because Eq. (1.2.5) must be solved n times. If the inverse matrix \mathbf{K}^{-1} is not tracked and is recalculated for each design variable, the amount of calculation required is similar to that when the finite-difference method is used to compute the cross-sectional derivative. \square

In Remark 1.2.3 (2), the finite-difference method was used as a comparison. The method for calculating the cross-sectional derivative by the finite-difference method is described as follows. Let $\mathbf{g}_i = (g_{ij})_{j \in \{1, \dots, n\}}$. Using the solutions to state equations corresponding to \mathbf{a} and $\mathbf{a} + (0, \dots, 0, b_j, 0, \dots, 0)^\top$, we calculate g_{ij} as follows:

$$g_{ij} = \frac{\tilde{f}_i\left(\mathbf{a} + (0, \dots, 0, b_j, 0, \dots, 0)^\top\right) - \tilde{f}_i(\mathbf{a})}{b_j}. \quad (1.2.6)$$

Thus, finding \mathbf{g}_i through this method requires solving the state equation n times.

1.2.2 The Adjoint Variable Method

Next, let us look at the adjoint variable method for calculating \mathbf{g}_i . Let the Lagrange function with respect to $f_i(\mathbf{a}, \mathbf{u})$ be

$$\begin{aligned} \mathcal{L}_i(\mathbf{a}, \mathbf{u}, \mathbf{v}_i) &= f_i(\mathbf{a}, \mathbf{u}) + \mathcal{L}_S(\mathbf{a}, \mathbf{u}, \mathbf{v}_i) \\ &= f_i(\mathbf{a}, \mathbf{u}) - \mathbf{v}_i \cdot (\mathbf{K}(\mathbf{a})\mathbf{u} - \mathbf{p}), \end{aligned} \quad (1.2.7)$$

where we have supposed that \mathcal{L}_S is defined by Eq. (1.2.2). Here, $\mathbf{v}_i \in \mathbb{R}^n$ is the adjoint variable (Lagrange multiplier) with respect to the state equation. The derivative of \mathcal{L}_i with respect to an arbitrary variation $(\mathbf{b}, \hat{\mathbf{u}}, \hat{\mathbf{v}}_i) \in X \times U \times U$ of

$(\mathbf{a}, \mathbf{u}, \mathbf{v}_i)$ is expressed as

$$\begin{aligned}\mathcal{L}'_i(\mathbf{a}, \mathbf{u}, \mathbf{v}_i) [\mathbf{b}, \hat{\mathbf{u}}, \hat{\mathbf{v}}_i] \\ = \mathcal{L}_{ia}(\mathbf{a}, \mathbf{u}, \mathbf{v}_i) [\mathbf{b}] + \mathcal{L}_{iu}(\mathbf{a}, \mathbf{u}, \mathbf{v}_i) [\hat{\mathbf{u}}] + \mathcal{L}_{iv_i}(\mathbf{a}, \mathbf{u}, \mathbf{v}_i) [\hat{\mathbf{v}}_i].\end{aligned}\quad (1.2.8)$$

The third term on the right-hand side of Eq. (1.2.8) satisfies

$$\mathcal{L}_{iv_i}(\mathbf{a}, \mathbf{u}, \mathbf{v}_i) [\hat{\mathbf{v}}_i] = \mathcal{L}_S(\mathbf{a}, \mathbf{u}, \hat{\mathbf{v}}_i). \quad (1.2.9)$$

Equation (1.2.9) is a Lagrange function with respect to the state determination problem (Problem 1.2.1). It is zero when \mathbf{u} solves the state determination problem.

Moreover, computing the second term on the right-hand side of Eq. (1.2.8) yields

$$\begin{aligned}\mathcal{L}_{iu}(\mathbf{a}, \mathbf{u}, \mathbf{v}_i) [\hat{\mathbf{u}}] &= f_{iu}(\mathbf{a}, \mathbf{u}) [\hat{\mathbf{u}}] + \mathcal{L}_{Su}(\mathbf{a}, \mathbf{u}, \mathbf{v}_i) [\hat{\mathbf{u}}] \\ &= f_{iu}(\mathbf{a}, \mathbf{u}) \cdot \hat{\mathbf{u}} - \mathbf{v}_i \cdot (\mathbf{K}(\mathbf{a}) \hat{\mathbf{u}}) \\ &= -\hat{\mathbf{u}} \cdot \left(\mathbf{K}^\top(\mathbf{a}) \mathbf{v}_i - \frac{\partial f_i}{\partial \mathbf{u}}(\mathbf{a}, \mathbf{u}) \right).\end{aligned}\quad (1.2.10)$$

Here, if \mathbf{v}_i can be determined so that Eq. (1.2.10) is zero for arbitrary $\hat{\mathbf{u}} \in U$, then the second term on the right-hand side of Eq. (1.2.8) vanishes. This condition is equivalent to setting \mathbf{v}_i to solve the following adjoint problem.

Problem 1.2.4 (The Adjoint Problem with Respect to f_i) Let $\mathbf{K}(\mathbf{a})$ and f_i be as in Problem 1.2.1. Find $\mathbf{v}_i \in U$ satisfying

$$\mathbf{K}^\top(\mathbf{a}) \mathbf{v}_i = f_{iu}(\mathbf{a}, \mathbf{u}). \quad \square$$

If \mathbf{u} and \mathbf{v}_i are solutions of Problem 1.2.1 and Problem 1.2.4, respectively, then we have

$$\begin{aligned}\mathcal{L}_{ia}(\mathbf{a}, \mathbf{u}, \mathbf{v}_i) [\mathbf{b}] \\ &= f_{ia}(\mathbf{a}, \mathbf{u}) \cdot \mathbf{b} - \mathbf{v}_i \cdot \left\{ \left(\frac{\partial \mathbf{K}(\mathbf{a})}{\partial a_1} \mathbf{u} \cdots \frac{\partial \mathbf{K}(\mathbf{a})}{\partial a_n} \mathbf{u} \right) \mathbf{b} \right\} \\ &= \left\{ f_{ia}(\mathbf{a}, \mathbf{u}) - \left(\frac{\partial \mathbf{K}(\mathbf{a})}{\partial a_1} \mathbf{u} \cdots \frac{\partial \mathbf{K}(\mathbf{a})}{\partial a_n} \mathbf{u} \right)^\top \mathbf{v}_i \right\} \cdot \mathbf{b} \\ &= \mathbf{g}_i \cdot \mathbf{b}.\end{aligned}$$

This result agrees with the formula obtained from the direct differentiation method (see Sect. 2.6.5).

The discussion above leads us to the following observations regarding the adjoint variable method.

Remark 1.2.5 (Properties of the Adjoint Variable Method)

In comparison with the method of direct differentiation, the adjoint variable method displays the following characteristics:

- (1) When the number of cost functions is large ($m \gg 1$), the adjoint variable method is ineffective because the number of adjoint problems is the same as the number of cost functions, $m + 1$.
- (2) When the number of design variables is large ($n \gg 1$), the adjoint variable method is effective because the number of adjoint problems, $m + 1$, does not depend on the number of design variables, n .
- (3) The number of variables in an adjoint problem is the same as the number of variables in the state equation, n . This indicates the fact that the adjoint variable method is applicable even when the state variable is a function of time or defined over a domain (in such a case, the linear space for the state variable becomes infinite-dimensional).
- (4) If a self-adjoint relationship is satisfied, then there is no need to explicitly solve the adjoint problem. \square

Beginning in Chap. 5, the state equation will be assumed to be a partial differential equation given by a boundary value problem. In this case, the state variable becomes a function defined in a domain of $d \in \{2, 3\}$ -dimensional space. Remark 1.2.5 (3) above indicates that the adjoint variable method is indispensable when constructing cost function derivatives with respect to design variables while satisfying the state equation in shape and topology optimization problems. However, in cases wherein a relation such as Eq. (1.2.5) could be obtained by a simple partial differential equation (for example, Eq. (8.5.17) in Chap. 8) under appropriate assumptions, then the direct differentiation method is effective in shape and topology optimization problems.

1.3 An Optimal Design Problem of a Branched One-Dimensional Stokes Flow Field

So far we have looked at what optimal design problems are by considering one-dimensional linear elastic bodies. Let us now continue our investigation by taking a look at how a similar optimal design problem can be constructed when the design target is changed to a flow field.

Consider a viscous flow field within a circular tube such as that shown in Fig. 1.6. This problem hints at Murray's law. By minimizing the sum of a blood flow transportation cost under a volume constraint, Murray showed that fluid flow through blood vessel cross-sections is proportional to the cube of the vessel radius [126]. Murray's analysis did not include a branched tube, but here we will take the cross-sectional areas as the design variables to see if the relationship still holds.

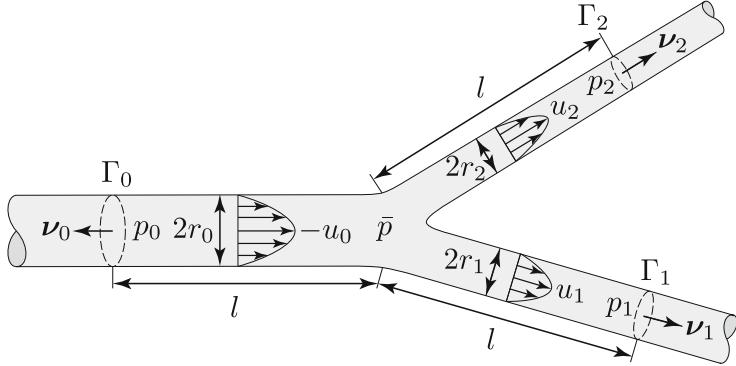


Fig. 1.6 A branched 1D Stokes flow field

Using the same cost function, Murray also derived a branch law for branch angles. The interested reader is referred to [125].

1.3.1 State Determination Problem

In Fig. 1.6, r_0 , r_1 , and r_2 denote the radii of their respective circular tubes, p_0 , p_1 , and p_2 signify the pressure at the inflow cross-section Γ_0 and the outflow cross-sections Γ_1 and Γ_2 , respectively. The pressure at the branched cross-section is denoted by \bar{p} , l represents the length, and μ is the viscosity coefficient. The flow velocity at a radius r within a circular cross-section of radius r_i , $i \in \{0, 1, 2\}$, is assumed to be given by the Hagen-Poiseuille flow:

$$u_{Hi}(r) = - \frac{p_i - \bar{p}}{4\mu l} (r_i^2 - r^2), \quad (1.3.1)$$

which is derived from the solution of a stationary Stokes equation with respect to a cylindrical boundary. Note that the flow velocities within the three tubes are taken to be positive in the outward normal directions ν_0 , ν_1 and ν_2 of the cross-sections Γ_0 , Γ_1 and Γ_2 , respectively. Meanwhile, due to the fact that $u_{H0}(r)$ flows inward from Γ_0 in the opposite direction of ν_0 , its flow velocity is negative. Here, let the volume of fluid flow per unit time through the tube of radius r_i be

$$u_i = \int_0^{r_i} u_{Hi}(r) 2\pi r \, dr = \frac{\bar{p} - p_i}{8\pi\mu l} (\pi r_i^2)^2 = (\bar{p} - p_i) a_i^2, \quad (1.3.2)$$

where

$$\mathbf{a} = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \frac{1}{\sqrt{8\pi\mu l}} \begin{pmatrix} \pi r_0^2 \\ \pi r_1^2 \\ \pi r_2^2 \end{pmatrix}. \quad (1.3.3)$$

According to Eq. (1.3.3), the cross-sectional area of Γ_i is $\sqrt{8\pi\mu l}a_i$. However, for the sake of simplicity a_i will be referred to as the cross-section and $\mathbf{a} \in X \in \mathbb{R}^3$ will be used as the design variable. Moreover, the continuity equation states that

$$u_0 + u_1 + u_2 = 0. \quad (1.3.4)$$

Substituting Eq. (1.3.2) into Eq. (1.3.4) yields

$$\bar{p} = \frac{p_0 a_0^2 + p_1 a_1^2 + p_2 a_2^2}{a_0^2 + a_1^2 + a_2^2}. \quad (1.3.5)$$

If Eq. (1.3.5) is substituted into Eq. (1.3.2), and \bar{p} is eliminated, then we obtain

$$\begin{aligned} & \frac{1}{a_0^2 + a_1^2 + a_2^2} \begin{pmatrix} a_0^2 (a_1^2 + a_2^2) & -a_0^2 a_1^2 & -a_0^2 a_2^2 \\ -a_0^2 a_1^2 & a_1^2 (a_0^2 + a_2^2) & -a_1^2 a_2^2 \\ -a_0^2 a_2^2 & -a_1^2 a_2^2 & a_2^2 (a_0^2 + a_1^2) \end{pmatrix} \begin{pmatrix} p_0 \\ p_1 \\ p_2 \end{pmatrix} \\ &= - \begin{pmatrix} u_0 \\ u_1 \\ u_2 \end{pmatrix}. \end{aligned} \quad (1.3.6)$$

In this section, we will assume that the volume of fluid flow per unit time $\mathbf{u} = (u_1, u_2)^\top \in \mathbb{R}^2$ is known and that u_0 is given by Eq. (1.3.4). The values relating to the flow velocity are assumed to be known because, as will be shown in Chap. 5, the existence of solutions to the Stokes problem is guaranteed when the flow velocity is given along the entire boundary (Theorem 5.6.3). However, the matrix of coefficients in this equation is singular because the equation holds regardless of the selected values for the average pressure (uncertainty with respect to the constant term). We therefore set $p_0 = 0$ for convenience. When the flow volume per unit time \mathbf{u} and the design variable \mathbf{a} are given, the pressure $\mathbf{p} = (p_1, p_2)^\top \in P = \mathbb{R}^2$ is then uniquely determined by

$$\frac{1}{a_0^2 + a_1^2 + a_2^2} \begin{pmatrix} a_1^2 (a_0^2 + a_2^2) & -a_1^2 a_2^2 \\ -a_1^2 a_2^2 & a_2^2 (a_0^2 + a_1^2) \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = - \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}. \quad (1.3.7)$$

Hence, $\mathbf{p} = (p_1, p_2)^\top \in P = \mathbb{R}^2$ is used as the state variable in this section and the state determination problem is defined as follows.

Problem 1.3.1 (1D Branched Stokes Flow Field) Consider the one-dimensional Stokes flow field of Fig. 1.6. Let $\mathbf{a} \in \mathbb{R}^3$ and $\mathbf{u} \in \mathbb{R}^2$ be given. Find $\mathbf{p} \in \mathbb{R}^2$ satisfying

$$\mathbf{A}(\mathbf{a}) \mathbf{p} = -\mathbf{u}. \quad (1.3.8)$$

Here, Eq. (1.3.8) expresses Eq. (1.3.7) in vector notation. \square

The solution of Eq. (1.3.8) is

$$\begin{aligned} \mathbf{p} &= -\mathbf{A}^{-1}(\mathbf{a}) \mathbf{u} \\ &= -\begin{pmatrix} \frac{1}{a_0^2} + \frac{1}{a_1^2} & \frac{1}{a_0^2} \\ \frac{1}{a_0^2} & \frac{1}{a_0^2} + \frac{1}{a_2^2} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \\ &= -\begin{pmatrix} \frac{u_1}{a_1^2} + \frac{u_1 + u_2}{a_0^2} \\ \frac{u_2}{a_2^2} + \frac{u_1 + u_2}{a_0^2} \end{pmatrix} = -\begin{pmatrix} \frac{u_1}{a_1^2} - \frac{u_0}{a_0^2} \\ \frac{u_2}{a_2^2} - \frac{u_0}{a_0^2} \end{pmatrix}. \end{aligned} \quad (1.3.9)$$

For later use, we define a Lagrange function with respect to Problem 1.3.1:

$$\mathcal{L}_S(\mathbf{a}, \mathbf{p}, \mathbf{q}) = \mathbf{q} \cdot (\mathbf{A}(\mathbf{a}) \mathbf{p} + \mathbf{u}), \quad (1.3.10)$$

where \mathbf{p} is not necessarily the solution of Problem 1.3.1 and $\mathbf{q} = (q_1, q_2)^\top \in \mathbb{R}^2$ is introduced as a Lagrange multiplier. Comparing Eq. (1.3.10) with Eq. (1.1.12), the sign convention for the Lagrangian looks different. This sign was decided to obtain the self-adjoint relationship of Eq. (1.3.18) together with the sign of \mathcal{L}_S in \mathcal{L}_0 defined later in Eq. (1.3.14). If the opposite sign was used for the right-hand side of Eq. (1.3.10), then the self-adjoint relationship of Eq. (1.3.18) holds with the opposite sign. This difference comes from the difference of the objective functions. The mean compliance represents the magnitude of displacement, while the mean flow resistance represents the negative value of the magnitude of flow velocity. Problem 1.3.1 is equivalent to finding \mathbf{p} satisfying

$$\mathcal{L}_S(\mathbf{a}, \mathbf{p}, \mathbf{q}) = 0,$$

for all $\mathbf{q} \in \mathbb{R}^2$.

1.3.2 An Optimal Design Problem

Having defined the state determination problem, we now establish a cost function using the design variable \mathbf{a} and the state variable \mathbf{p} .

We want to construct an objective function that measures flow resistance. According to the law of conservation of energy, the energy lost through viscosity per unit time inside the viscous flow field is equal to the negative value of the power (energy per unit time) integrated along the boundary. Now, let the objective function be

$$f_0 = -(p_0 u_0 + p_1 u_1 + p_2 u_2) = -\mathbf{p} \cdot \mathbf{u}, \quad (1.3.11)$$

where we have used the fact that $p_0 = 0$. In Eq. (1.3.11), assuming \mathbf{u} is given, the minimization of f_0 means the maximization of \mathbf{p} at the out flow boundaries. It means the minimization of pressure loss. Then, this function corresponds to values often referred to as dissipation energy or power loss. However, in Chaps. 8 and 9, an extension of this definition (referred to as the mean flow resistance) will be used as a cost function to measure flow resistance in a Stokes flow field. The reason for this terminology is because, in that scenario, the quantity does not represent dissipation energy. For this reason, f_0 of Eq. (1.3.11) will also be referred to as the mean flow resistance in a one-dimensional branched Stokes flow field.

The volume constraint function is taken as

$$f_1(\mathbf{a}) = l(a_0 + a_1 + a_2) - c_1, \quad (1.3.12)$$

where c_1 is a positive constant.

Having defined these cost functions, the optimization problem for the one-dimensional branched Stokes flow field is defined as follows. We take $X = \mathbb{R}^3$ as the linear space for the design variable \mathbf{a} and, with respect to a constant vector $\mathbf{a}_0 > \mathbf{0}_{\mathbb{R}^3}$, we set

$$\mathcal{D} = \{\mathbf{a} \in X \mid \mathbf{a} \geq \mathbf{a}_0\}. \quad (1.3.13)$$

Furthermore, $P = \mathbb{R}^2$ denotes the linear space for the respective state variables, \mathbf{p} .

Problem 1.3.2 (Mean Flow Resistance Minimization) Let $X = \mathbb{R}^3$, $P = \mathbb{R}^2$, and \mathcal{D} be given by Eq. (1.3.13). Furthermore, let $f_0(\mathbf{p})$ and $f_1(\mathbf{a})$ be given by Eqs. (1.3.11) and (1.3.12), respectively. Find \mathbf{a} satisfying

$$\min_{(\mathbf{a}, \mathbf{p}) \in \mathcal{D} \times P} \{f_0(\mathbf{p}) \mid f_1(\mathbf{a}) \leq 0, \text{ Problem 1.3.1}\}. \quad \square$$

1.3.3 Cross-Sectional Derivatives

The derivative $\tilde{f}'_0(\mathbf{a})[\mathbf{b}] = f'_0(\mathbf{p}(\mathbf{a}))[\mathbf{b}] = \mathbf{g}_0 \cdot \mathbf{b}$ of f_0 with respect to a variation \mathbf{b} of \mathbf{a} is called the cross-sectional derivative and \mathbf{g}_0 is referred to as the cross-sectional-area gradient. Let us now find \mathbf{g}_0 and the Hesse matrix \mathbf{H}_0 of f_0 using the adjoint variable method.

The Lagrange function with respect to f_0 is taken as

$$\begin{aligned}\mathcal{L}_0(\mathbf{a}, \mathbf{p}, \mathbf{q}_0) &= f_0(\mathbf{p}) - \mathcal{L}_S(\mathbf{a}, \mathbf{p}, \mathbf{q}_0) \\ &= -\mathbf{p} \cdot \mathbf{u} - \mathbf{q}_0 \cdot (\mathbf{A}(\mathbf{a})\mathbf{p} + \mathbf{u}),\end{aligned}\quad (1.3.14)$$

where $\mathbf{q}_0 \in P$ is the adjoint variable (Lagrange multiplier) with respect to the state equation (prepared for f_0). The derivative of \mathcal{L}_0 with respect to an arbitrary variation $(\mathbf{b}, \hat{\mathbf{p}}, \hat{\mathbf{q}}_0) \in X \times P \times P$ of $(\mathbf{a}, \mathbf{p}, \mathbf{q}_0)$ is

$$\begin{aligned}\mathcal{L}'_0(\mathbf{a}, \mathbf{p}, \mathbf{q}_0)[\mathbf{b}, \hat{\mathbf{p}}, \hat{\mathbf{q}}_0] &= \mathcal{L}_{0a}(\mathbf{a}, \mathbf{p}, \mathbf{q}_0)[\mathbf{b}] + \mathcal{L}_{0p}(\mathbf{a}, \mathbf{p}, \mathbf{q}_0)[\hat{\mathbf{p}}] + \mathcal{L}_{0q_0}(\mathbf{a}, \mathbf{p}, \mathbf{q}_0)[\hat{\mathbf{q}}_0].\end{aligned}\quad (1.3.15)$$

The third term on the right-hand side of Eq. (1.3.15) satisfies

$$\mathcal{L}_{0q_0}(\mathbf{a}, \mathbf{p}, \mathbf{q}_0)[\hat{\mathbf{q}}_0] = -\mathcal{L}_S(\mathbf{a}, \mathbf{p}, \hat{\mathbf{q}}_0). \quad (1.3.16)$$

This term is zero if \mathbf{p} solves the state determination problem.

The second term on the right-hand side of Eq. (1.3.15) can be calculated:

$$\begin{aligned}\mathcal{L}_{0p}(\mathbf{a}, \mathbf{p}, \mathbf{q}_0)[\hat{\mathbf{p}}] &= f_{0p}(\mathbf{a}, \mathbf{p})[\hat{\mathbf{p}}] - \mathcal{L}_{Sp}(\mathbf{a}, \mathbf{p}, \mathbf{q}_0)[\hat{\mathbf{p}}] \\ &= -\mathcal{L}_S(\mathbf{a}, \mathbf{q}_0, \hat{\mathbf{p}}).\end{aligned}\quad (1.3.17)$$

This term can also be made to take the value zero, provided the self-adjoint relationship holds:

$$\mathbf{q}_0 = \mathbf{p}. \quad (1.3.18)$$

Furthermore, direct calculation shows that the first term on the right-hand side of Eq. (1.3.15) can be expressed as

$$\begin{aligned}\mathcal{L}_{0a}(\mathbf{a}, \mathbf{p}, \mathbf{q}_0)[\mathbf{b}] &= -\frac{1}{(a_0^2 + a_1^2 + a_2^2)^2}\end{aligned}$$

$$\begin{aligned}
& \times \begin{pmatrix} 2a_0(a_1^2 p_1 + a_2^2 p_2)(a_1^2 q_{01} + a_2^2 q_{02}) \\ 2a_1 \{a_0^2 p_1 + a_2^2 (p_1 - p_2)\} \{a_0^2 q_{01} + a_2^2 (q_{01} - q_{02})\} \\ 2a_2 \{a_0^2 p_2 + a_1^2 (p_2 - p_1)\} \{a_0^2 q_{02} + a_1^2 (q_{02} - q_{01})\} \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} \\
& = -2 \begin{pmatrix} u_0^2/a_0^3 \\ u_1^2/a_1^3 \\ u_2^2/a_2^3 \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \mathbf{g}_0 \cdot \mathbf{b}. \tag{1.3.19}
\end{aligned}$$

Here, the self-adjoint relationship has been used along with the fact that \mathbf{p} is a solution of the state determination problem.

It can also be easily seen that

$$f'_1(\mathbf{a})[\mathbf{b}] = l \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \mathbf{g}_1 \cdot \mathbf{b}. \tag{1.3.20}$$

Furthermore, the Hesse matrix \mathbf{H}_0 of the mean flow resistance $\tilde{f}_0(\mathbf{a}) = f_0(\mathbf{a}, \mathbf{p}(\mathbf{a}))$ can be obtained as follows. As described in Sect. 1.1.6, we use the adjoint variable method. The second-order derivative of the Lagrange function \mathcal{L}_0 with respect to arbitrary variations $(\mathbf{b}_1, \hat{\mathbf{p}}_1) \in X \times P$ and $(\mathbf{b}_2, \hat{\mathbf{p}}_2) \in X \times P$ of the design and state variables (\mathbf{a}, \mathbf{p}) is computed as follows:

$$\begin{aligned}
& \mathcal{L}_{0(\mathbf{a}, \mathbf{p}), (\mathbf{a}, \mathbf{p})}(\mathbf{a}, \mathbf{p}, \mathbf{q}_0)[(\mathbf{b}_1, \hat{\mathbf{p}}_1), (\mathbf{b}_2, \hat{\mathbf{p}}_2)] \\
& = (\mathcal{L}_{0\mathbf{a}}(\mathbf{a}, \mathbf{p}, \mathbf{q}_0)[\mathbf{b}_1] + \mathcal{L}_{0\mathbf{p}}(\mathbf{a}, \mathbf{p}, \mathbf{q}_0)[\hat{\mathbf{p}}_1])_{\mathbf{a}}[\mathbf{b}_2] \\
& \quad + (\mathcal{L}_{0\mathbf{a}}(\mathbf{a}, \mathbf{p}, \mathbf{q}_0)[\mathbf{b}_1] + \mathcal{L}_{0\mathbf{p}}(\mathbf{a}, \mathbf{p}, \mathbf{q}_0)[\hat{\mathbf{p}}_1])_{\mathbf{u}}[\hat{\mathbf{p}}_2] \\
& = \begin{pmatrix} \mathbf{b}_2 \\ \hat{\mathbf{p}}_2 \end{pmatrix} \cdot \begin{pmatrix} (\mathcal{L}_{0\mathbf{aa}} \mathcal{L}_{0\mathbf{ap}}) & (\mathbf{b}_1) \\ (\mathcal{L}_{0\mathbf{pa}} \mathcal{L}_{0\mathbf{pp}}) & (\hat{\mathbf{p}}_1) \end{pmatrix}. \tag{1.3.21}
\end{aligned}$$

Here, $\hat{\mathbf{p}}_1$ and $\hat{\mathbf{p}}_2$ are replaced by variations satisfying the state determination problem. By taking the partial derivative of Eq. (1.3.8) with respect to a_i for $i \in \{1, 2\}$, we obtain

$$\frac{\partial \mathbf{A}}{\partial a_i} \mathbf{p} + \mathbf{A} \frac{\partial \mathbf{p}}{\partial a_i} = \mathbf{0}_{\mathbb{R}^2}. \tag{1.3.22}$$

Solving for $\partial \mathbf{p} / \partial a_i$, we get

$$\frac{\partial \mathbf{p}}{\partial a_i} = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial a_i} \mathbf{p}, \tag{1.3.23}$$

and set

$$\hat{\mathbf{p}}(\mathbf{a})[\mathbf{b}] = \frac{\partial \mathbf{p}}{\partial \mathbf{a}^\top} \mathbf{b} = \begin{pmatrix} \partial p_1 / \partial a_0 & \partial p_1 / \partial a_1 & \partial p_1 / \partial a_2 \\ \partial p_2 / \partial a_0 & \partial p_2 / \partial a_1 & \partial p_2 / \partial a_2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}. \quad (1.3.24)$$

Substituting Eq. (1.3.24) into Eq. (1.3.21), we obtain

$$\mathcal{L}_{0(\mathbf{a}, \mathbf{p}), (\mathbf{a}, \mathbf{p})}(\mathbf{a}, \mathbf{p}, \mathbf{q}_0) [(\mathbf{b}_1, \hat{\mathbf{p}}(\mathbf{a})[\mathbf{b}_1]), (\mathbf{b}_2, \hat{\mathbf{p}}(\mathbf{a})[\mathbf{b}_2])] = \mathbf{b}_1 \cdot (\mathbf{H}_0 \mathbf{b}_2), \quad (1.3.25)$$

and we see that, if the self-adjoint relationship is used along with the fact that \mathbf{p} is a solution of the state determination problem, the Hesse matrix of \tilde{f}_0 is expressed as

$$\begin{aligned} \mathbf{H}_0 &= \mathcal{L}_{0aa} + \mathcal{L}_{0ap} \frac{\partial \mathbf{p}}{\partial \mathbf{a}^\top} + \left(\mathcal{L}_{0ap} \frac{\partial \mathbf{p}}{\partial \mathbf{a}^\top} \right)^\top \\ &= 6 \begin{pmatrix} u_0^2/a_0^4 & 0 & 0 \\ 0 & u_1^2/a_1^4 & 0 \\ 0 & 0 & u_2^2/a_2^4 \end{pmatrix}. \end{aligned} \quad (1.3.26)$$

This formula matches the result obtained through partial differentiation of \mathbf{g}_0 in Eq. (1.3.19) with respect to \mathbf{a} . This relationship holds true because the state variable \mathbf{p} is not included in \mathbf{g}_0 of Eq. (1.3.19). In this way, we observe that \mathbf{H}_0 is positive definite.

Let us, in addition, obtain the Hesse matrix of \tilde{f}_0 by the Lagrange multiplier method. The Lagrange function for $\mathbf{g}_0 \cdot \mathbf{b}_1$ can be defined as

$$\begin{aligned} \mathcal{L}_{10}(\mathbf{a}, \mathbf{p}, \mathbf{r}_0) &= \mathbf{g}_0(\mathbf{p}) \cdot \mathbf{b}_1 - \mathcal{L}_S(\mathbf{a}, \mathbf{p}, \mathbf{r}_0) \\ &= \mathbf{g}_0(\mathbf{p}) \cdot \mathbf{b}_1 - \mathbf{r}_0 \cdot (\mathbf{A}(\mathbf{a}) \mathbf{p} + \mathbf{u}), \end{aligned} \quad (1.3.27)$$

where $\mathbf{g}_0(\mathbf{p})$ is defined in the second equation of Eq. (1.3.19) substituting Eq. (1.3.18), \mathcal{L}_S in Eq. (1.3.10). $\mathbf{r}_0 = (r_{01}, r_{02})^\top \in P = \mathbb{R}^2$ is the adjoint variable corresponding to \mathbf{p} in $\mathbf{g}_0(\mathbf{p})$ satisfying the state determination problem. \mathbf{b}_1 was assumed to be a constant vector in \mathcal{L}_{10} .

With respect to arbitrary variations $(\mathbf{b}_2, \hat{\mathbf{p}}, \hat{\mathbf{r}}_0) \in X \times P^2$ of $(\mathbf{a}, \mathbf{p}, \mathbf{r}_0)$, the derivative of \mathcal{L}_{10} is written as

$$\begin{aligned} \mathcal{L}'_{10}(\mathbf{a}, \mathbf{p}, \mathbf{r}_0) [\mathbf{b}_2, \hat{\mathbf{p}}, \hat{\mathbf{r}}_0] &= \mathcal{L}_{10a}(\mathbf{a}, \mathbf{p}, \mathbf{r}_0) [\mathbf{b}_2] + \mathcal{L}_{10p}(\mathbf{a}, \mathbf{p}, \mathbf{r}_0) [\hat{\mathbf{p}}] + \mathcal{L}_{10r_0}(\mathbf{a}, \mathbf{p}, \mathbf{r}_0) [\hat{\mathbf{r}}_0]. \end{aligned} \quad (1.3.28)$$

The third term on the right-hand side of Eq. (1.3.28) vanishes if \mathbf{p} is the solution of the state determination problem.

The second term on the right-hand side of Eq. (1.3.28) can be written as

$$\begin{aligned}\mathcal{L}_{10p}(\mathbf{a}, \mathbf{p}, \mathbf{r}_0)[\hat{\mathbf{p}}] &= \mathbf{g}_{0p}^\top(\mathbf{p})[\hat{\mathbf{p}}] \cdot \mathbf{b}_1 - \mathcal{L}_{Sp}(\mathbf{a}, \mathbf{p}, \mathbf{r}_0)[\hat{\mathbf{p}}] \\ &= \hat{\mathbf{p}} \cdot \mathbf{w} - \mathbf{r}_0 \cdot (\mathbf{A}(\mathbf{a}) \hat{\mathbf{p}}) \\ &= \hat{\mathbf{p}} \cdot (\mathbf{w} - \mathbf{A}^\top(\mathbf{a}) \mathbf{r}_0),\end{aligned}\quad (1.3.29)$$

where

$$\mathbf{w} = \mathbf{g}_{0p}^\top(\mathbf{p}) \mathbf{b}_1 = \begin{pmatrix} \frac{\partial g_{01}}{\partial p_1} & \frac{\partial g_{02}}{\partial p_1} & \frac{\partial g_{03}}{\partial p_1} \\ \frac{\partial g_{01}}{\partial p_2} & \frac{\partial g_{02}}{\partial p_2} & \frac{\partial g_{03}}{\partial p_2} \end{pmatrix} \begin{pmatrix} b_{11} \\ b_{12} \\ b_{13} \end{pmatrix}. \quad (1.3.30)$$

Here, the condition that Eq. (1.3.29) is zero for arbitrary $\hat{\mathbf{p}} \in P$ is equivalent to setting \mathbf{r}_0 to be the solution of the following adjoint problem.

Problem 1.3.3 (Adjoint Problem of \mathbf{r}_0 with Respect to $g_0(\mathbf{p}) \cdot \mathbf{b}_1$) Let $\mathbf{A}(\mathbf{a})$ be as in Problem 1.3.1, and \mathbf{w} be given by Eq. (1.3.30). Find $\mathbf{r}_0 \in P$ satisfying

$$\mathbf{A}^\top(\mathbf{a}) \mathbf{r}_0 = \mathbf{w}.$$

□

The solution of Problem 1.3.3 is

$$\mathbf{r}_0 = \left(\mathbf{A}^\top(\mathbf{a}) \right)^{-1} \mathbf{g}_{0p}^\top(\mathbf{p}) \mathbf{b}_1. \quad (1.3.31)$$

Finally, the first term on the right-hand side of Eq. (1.3.28) becomes

$$\begin{aligned}\mathcal{L}_{10a}(\mathbf{a}, \mathbf{p}, \mathbf{r}_0)[\mathbf{b}_2] &= - \left\{ \mathbf{r}_0 \cdot \left(\frac{\partial \mathbf{A}(\mathbf{a})}{\partial a_1} \mathbf{p} \frac{\partial \mathbf{A}(\mathbf{a})}{\partial a_2} \mathbf{p} \frac{\partial \mathbf{A}(\mathbf{a})}{\partial a_3} \mathbf{p} \right) \right\} \mathbf{b}_2.\end{aligned}\quad (1.3.32)$$

Here, substituting Eq. (1.3.31) into Eq. (1.3.32), we have the relation

$$\mathcal{L}_{10a}(\mathbf{a}, \mathbf{p}, \mathbf{r}_0(\mathbf{b}_1))[\mathbf{b}_2] = h_0(\mathbf{a})[\mathbf{b}_1, \mathbf{b}_2] = \mathbf{g}_{H0}(\mathbf{a}, \mathbf{b}_1) \cdot \mathbf{b}_2, \quad (1.3.33)$$

where the Hesse gradient of the mean flow resistance \mathbf{g}_{H0} is given by

$$\mathbf{g}_{H0}(\mathbf{a}, \mathbf{b}_1) = \mathcal{L}_{10a}(\mathbf{a}, \mathbf{p}, \mathbf{r}_0(\mathbf{b}_1)). \quad (1.3.34)$$

The above results show that the function $\tilde{f}_0(\mathbf{a})$, which is obtained by rewriting the mean flow resistance $f_0(\mathbf{p})$ as a function of \mathbf{a} only, is convex. As we will now

show, as in the case of the mean compliance problem, \mathbf{a} yields the minimum when it satisfies a set of KKT conditions.

1.3.4 Optimality Conditions

Let us again consider optimality using the KKT conditions. The Lagrange function with respect to the optimization problem (Problem 1.3.2) is set as

$$\mathcal{L}(\mathbf{a}, \lambda_1) = \tilde{f}_0(\mathbf{a}) + \lambda_1 f_1(\mathbf{a}),$$

where $\lambda_1 \in \mathbb{R}$ is a Lagrange multiplier with respect to $f_1(\mathbf{a}) \leq 0$. In this case, the KKT conditions Karush–Kuhn–Tucker conditions of Problem 1.3.2 are given by

$$\mathcal{L}_{\mathbf{a}}(\mathbf{a}, \lambda_1) = \mathbf{g}_0 + \lambda_1 \mathbf{g}_1 = \mathbf{0}_{\mathbb{R}^2}, \quad (1.3.35)$$

$$\mathcal{L}_{\lambda_1}(\mathbf{a}, \lambda_1) = f_1(\mathbf{a}) \leq 0, \quad (1.3.36)$$

$$\lambda_1 f_1(\mathbf{a}) = 0, \quad (1.3.37)$$

$$\lambda_1 \geq 0. \quad (1.3.38)$$

Equation (1.3.35) then becomes

$$-2 \begin{pmatrix} u_0^2/a_0^3 \\ u_1^2/a_1^3 \\ u_2^2/2a_1^3 \end{pmatrix} + \lambda_1 l \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

and the optimality condition with respect to the mean flow resistance minimization problem (Problem 1.3.2) is given by

$$2 \frac{u_0^2}{a_0^3 l} = 2 \frac{u_1^2}{a_1^3 l} = 2 \frac{u_2^2}{a_2^3 l} = \lambda_1. \quad (1.3.39)$$

This optimality condition shows that Murray's law holds. In fact, using $u_{Hi}(r)$ from Eq. (1.3.1) shows that the shear strain velocity and shear stress on the walls can be expressed as

$$\gamma_i = \left. \frac{du_{Hi}}{dr} \right|_{r=r_i} = -\frac{\bar{p} - p_i}{2\mu l} r_i = -\frac{u_i}{2\mu l a_i^2} r_i, \quad (1.3.40)$$

$$\tau_i = \mu \gamma_i = -\frac{u_i}{2l a_i^2} r_i, \quad (1.3.41)$$

respectively, for each $i \in \{0, 1, 2\}$. Using Eq. (1.3.3), we then obtain the following relation regarding the dissipation energy density:

$$\frac{1}{2} \tau_i \gamma_i = \frac{1}{\sqrt{8\mu l}} \frac{u_i^2}{a_i^3 l} = \frac{\sqrt{8\mu l}}{2} \lambda_1. \quad (1.3.42)$$

When the shear stresses are the same, this condition shows that the flow volume is proportional to the cube of the blood vessel radius (Murray's law).

1.3.5 Numerical Example

Let us now consider finding a minimizer through a specific exercise.

Exercise 1.3.4 (Mean Flow Resistance Minimization Problem) Let $a_0 = 1$ in Problem 1.3.2 (a_0 is not included in the design variables). Also, let $l = 1$, $c_1 = 2$, $\mathbf{u} = (1/3, 2/3)^\top$, and $\mathbf{a}_0 = (0.1, 0.1, 0.1)^\top$. Find the minimizer \mathbf{a} . \square

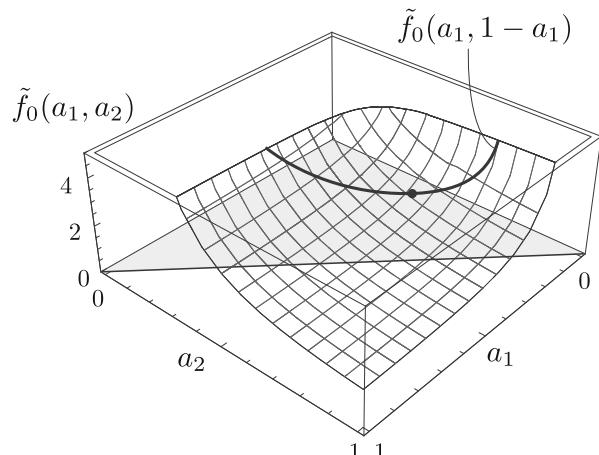
Answer If \mathbf{p} in Eq. (1.3.9) is substituted into f_0 (defined by Eq. (1.3.11)), then we obtain:

$$\tilde{f}_0(\mathbf{a}) = \frac{1}{9} \left(9 + \frac{1}{a_1} + \frac{4}{a_2} \right). \quad (1.3.43)$$

Figure 1.7 shows \tilde{f}_0 . From Eqs. (1.3.19) and (1.3.20), we obtain the cross-sectional-area derivative of \tilde{f}_0 and f_1 :

$$\mathbf{g}_0 = - \begin{pmatrix} 2/9a_1^3 \\ 8/9a_2^3 \end{pmatrix}, \quad \mathbf{g}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Fig. 1.7 Numerical example of mean flow resistance minimization problem



Here, if \mathbf{a} yields the minimum, then from Eq. (1.3.39) we get

$$\lambda_1 = \frac{2}{9a_1^3} = \frac{8}{9a_2^3}.$$

If \mathbf{a} is an element of \mathcal{D}° (defined by Eq. (1.3.13)), then λ_1 is positive and the inequality constraint with respect to f_1 holds as an equality (due to the complementarity condition). Upon substituting $a_2 = 1 - a_1$ into Eq. (1.3.43), and using the stationary condition with respect to variations in a_1 :

$$\frac{d}{da_1} \tilde{f}_0(a_1, 1 - a_1) = \frac{1}{(1 - a_1)^2} - \frac{4}{a_1^2} = 0,$$

we obtain

$$\begin{aligned} a_1 &= \frac{1}{5} \left(1 + 2^{4/3} - 2^{2/3} \right), \quad \frac{1}{5} \left\{ 1 - 2^{1/3} (1 - i\sqrt{3}) + 2^{-1/3} (1 + i\sqrt{3}) \right\}, \\ &\quad \frac{1}{5} \left\{ 1 - 2^{1/3} (1 + i\sqrt{3}) + 2^{-1/3} (1 - i\sqrt{3}) \right\} \\ &= 0.386488, \quad 0.106756 + 0.711395i, \quad 0.106756 - 0.711395i. \end{aligned}$$

Here i denotes the imaginary unit and we remark that a_2 can be obtained from $1 - a_1$. Using these results, $\mathbf{a} \in \mathcal{D}^\circ$ is then given by $\mathbf{a} = (0.386488, 0.613512)^\top$. Since \tilde{f}_0 and f_1 are convex-functions, Problem 1.3.2 is a convex-optimization problem and, from Theorem 2.7.9, the \mathbf{a} which satisfies the KKT condition yields the minimum. \square

1.4 Summary

In order to have a conceptional idea about how optimality conditions are sought, this chapter examined some examples of optimal design problems related to one-dimensional elastic bodies and the one-dimensional Stokes flow field. The key points are outlined below:

- (1) In addition to design variables that determine the system, optimal design problems include state variables which describe the system's desired state. The state equation which determines the state variables is an equality constraint. Cost functions can be defined as functions of the design and state variables (Sects. 1.1.2, 1.3.2).
- (2) When the cost function is given as a function of a state variable, its derivative with respect to a variation of the design variable needs to be obtained in a manner that satisfies the equality constraints of the state determination problem. The direct differentiation method (Sects. 1.1.5, 1.2.1), which uses the chain rule

of differentiation, and the adjoint variable method (Sects. 1.1.6, 1.2.2), which is based on the Lagrange multiplier method, are both conceivable methods for obtaining the derivative.

- The direct differentiation method is advantageous for problems involving multiple constraints (Remark 1.2.3).
- The adjoint variable method is advantageous for problems with multiple design variables (Remark 1.2.5).

Later in this book, partial differential equations (boundary value problems) are assumed to be the state equation, and the state variable becomes a function (an element of an infinite-dimensional space) defined in a $d \in \{2, 3\}$ -dimensional domain. The adjoint variable approach is essential in such settings.

- (3) In a mean compliance minimization problem (Problem 1.1.4), where the cross-sectional area of the one-dimensional elastic body was the design variable, an optimality condition stating that the strain energy density is uniform was obtained in Eq. (1.1.56) (Sect. 1.1.7).
- (4) In a mean flow resistance minimization problem (Problem 1.3.2), where the cross-sectional area of the branched one-dimensional Stokes flow field is the design variable, an optimality condition stating that the dissipative energy density is uniform was derived in Eq. (1.3.42) (Sect. 1.3.4).

There is a vast amount of literature regarding optimal design problems, and here we only mention a selection [6, 39, 61, 65, 181].

1.5 Practice Problems

1.1 Consider the problem of finding \mathbf{a} satisfying

$$\min_{(\mathbf{a}, \mathbf{u}) \in \mathcal{D} \times U} \left\{ f_0(\mathbf{u}) = u_2^2 \mid f_1(\mathbf{a}) \leq 0, \text{ Problem 1.1.3} \right\}.$$

This is Problem 1.1.4 with f_0 changed to u_2^2 . Let the adjoint variable with respect to f_0 be $\mathbf{v}_0 \in \mathbb{R}^2$. Derive the adjoint equation. Also, express \mathbf{g}_0 in terms of \mathbf{u} and \mathbf{v}_0 .

1.2 In Sect. 1.1.6, using the stationary conditions with respect to an arbitrary variation of \mathbf{u} and \mathbf{v}_0 of the Lagrange function in Problem 1.1.4, we derived the cross-sectional-area gradient \mathbf{g}_0 of f_0 . In this case, the self-adjoint relationship was used. In fact, if the self-adjoint relationship holds, then the cross-sectional-area gradient \mathbf{g}_0 can be obtained without using a Lagrange function. In particular, instead of using the mean compliance f_0 , the potential energy

$$\pi(\mathbf{a}, \mathbf{u}) = \frac{1}{2} \mathbf{u} \cdot (\mathbf{K}(\mathbf{a}) \mathbf{u}) - \mathbf{u} \cdot \mathbf{p}$$

can be used to consider the problem of finding (\mathbf{a}, \mathbf{u}) satisfying

$$\max_{\mathbf{a} \in \mathcal{D}} \min_{\mathbf{u} \in U} \pi(\mathbf{a}, \mathbf{u}).$$

When \mathbf{u} satisfies $\min_{\mathbf{u} \in \mathbb{R}^2} \pi$, show that the cross-sectional-area gradient of $-\pi$ is equal to $1/2$ of the \mathbf{g}_0 in Eq. (1.1.36). Note that π in this problem is the potential energy. This problem shows that the minimization of the mean compliance is equivalent to the maximization of the potential energy.

1.3 Consider Practice 1.1, where $l = 1$, $e_Y = 1$, $c_1 = 1$ and $\mathbf{p} = (1, 1)^\top$. Find the minimizer \mathbf{a} .

1.4 Consider the state determination problem in Problem 1.1.4. Here, when $\mathbf{p} = (1, -1)^\top$, the stress of a one-dimensional linear elastic body with a cross-section of a_1 is zero. The side constraint with respect to the cross-section a_1 thus activates at the optimal solution. The optimal solution in this case is $(a_1, a_2) = (a_{01}, (c_1/l) - a_{01})$. Derive the KKT conditions in this situation and find the Lagrange multiplier. Here, assume that \mathbf{g}_0 and \mathbf{g}_1 are given by Eqs. (1.1.28) and (1.1.17), respectively.

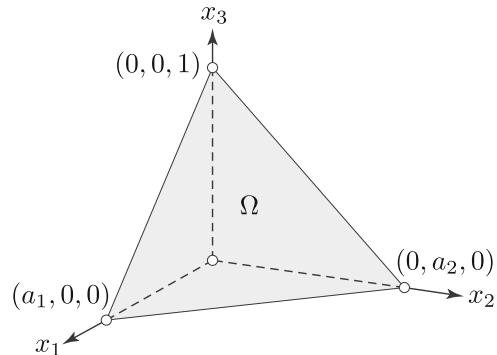
1.5 Consider the state determination problem of Problem 1.1.4 and assume that the design variable $\mathbf{a} \in \mathbb{R}^2$ is the length of one side of a square cross-section. Find the Hesse matrix \mathbf{H}_0 and the gradient \mathbf{g}_0 of f_0 , with respect to variation of \mathbf{a} .

1.6 Consider the tetrahedron Ω shown in Fig. 1.8. Let the design variable be the lengths of the sides of the base $\mathbf{a} = (a_1, a_2)^\top \in \mathbb{R}^2$, and the cost function f be the volume of Ω . Find the Hesse matrix \mathbf{H} and the gradient \mathbf{g} of f , with respect to variation of \mathbf{a} .

1.7 Give a concrete expression for $\mathbf{K}(\mathbf{a})$ in Eq. (1.2.1).

1.8 The self-adjoint relationship was also established with respect to f_0 in Problem 1.3.2. Thus, in a manner similar to Practice 1.2, the cross-sectional-area gradient \mathbf{g}_0 can be obtained without using a Lagrange function. Considering f_0 , let

Fig. 1.8 Tetrahedron Ω



us formally investigate the problem of finding (\mathbf{a}, \mathbf{p}) satisfying

$$\min_{\mathbf{a} \in \mathcal{D}} \max_{\mathbf{p} \in P} \pi(\mathbf{a}, \mathbf{p}),$$

where the potential energy of the dissipative system is set to be

$$\pi(\mathbf{a}, \mathbf{p}) = -\frac{1}{2} \mathbf{p} \cdot (\mathbf{A}(\mathbf{a}) \mathbf{p}) - \mathbf{u} \cdot \mathbf{p}.$$

When \mathbf{p} satisfying $\max_{\mathbf{p} \in \mathbb{R}^2} \pi$ is used, show that the cross-sectional-area gradient of π is the same as $1/2$ of \mathbf{g}_0 in Eq. (1.3.19). This problem shows that minimizing the mean flow resistance is the same as minimizing π when the potential energy of the dissipative system is formally set to π .

1.9 Consider a branched one-dimensional Stokes flow field such as the one shown in Fig. 1.9. The center of the inflow boundary Γ_0 is taken to be the origin, while $\alpha = (\alpha_1, -\alpha_2)^\top \in \mathbb{R}^2$ and $\beta = (\beta_1, \beta_2)^\top \in \mathbb{R}^2$ (with respect to four positive constants $\alpha_1, \alpha_2, \beta_1$ and β_2) are set as the coordinates of the central position of the outward flow boundaries, Γ_1 and Γ_2 . The radius of the tube is $\mathbf{r} = (r_0, r_1, r_2)^\top \in \mathbb{R}^3$. Assume that \mathbf{r}, α and β are known and that the sum of the volumes of the three tubes is minimized. Use the branch angles θ_1 and θ_2 to show that

$$r_0^2 = r_1^2 \cos \theta_1 + r_2^2 \cos \theta_2.$$

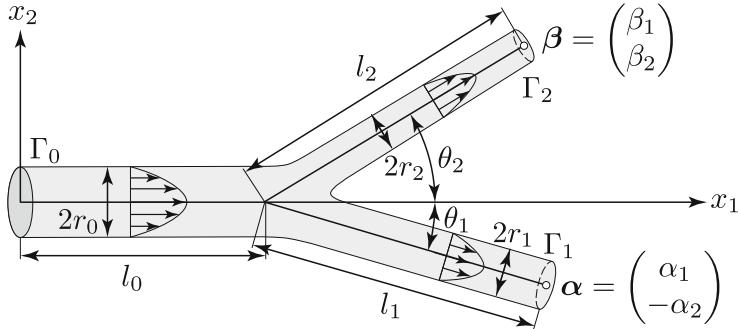


Fig. 1.9 The branch angle of a branched one-dimensional Stokes flow field

Chapter 2

Basics of Optimization Theory



Chapter 1 investigated explicit optimal design problems and illustrated different approaches for obtaining optimality conditions. Terminology and results utilized in optimization theory were also used. This chapter presents a systematic discussion of optimization theory.

The stage of this chapter is a finite-dimensional vector space. In other words, the linear spaces of the design and state variables are assumed to be of finite dimension. Later in this book, optimization problems will be constructed on function spaces and methods for their solution will be considered. However, many of the concepts and results of this chapter can also be used in the function space setting. In this sense, the content of this chapter can be seen as forming the foundation of this book. In other words, the extent of the reliability of this book is dependent on the content of this chapter. For this reason, although the details become somewhat abstracted, concepts will be summarized in the format of definitions and theorems. Here, in order to tie the abstract notions together with concrete ideas, our discussions will be interlaced with examples related to the simple spring systems that the reader is already familiar with.

2.1 Definition of Optimization Problems

Considering the problems presented in Chap. 1, let us first define the optimization problems that will be the target of this chapter's discussions. In Problem 1.2.2, the cross-section was set as $a \in X = \mathbb{R}^n$ and the displacement was $u \in U = \mathbb{R}^n$. These were referred to as the design and state variables, respectively, where X and U represented their linear spaces. Let us now define several optimization problems which include this one.

With respect to the linear spaces, we will refrain from using their symbols in ways which are not observed in other chapters. In Chap. 1, X and U represented

linear spaces containing the design and state variables. In Chap. 2, design and state variables are collectively referred to as design variables, and X represents the linear space of the design variables. That is to say, with respect to the variables of Chap. 1, we define $\mathbf{x} = (\mathbf{a}^\top, \mathbf{u}^\top)^\top \in X$. The definitions must be changed in this manner because design and state variables are not distinguished from each other in standard optimization problems (all variables are treated as design variables).

Next, let us define an optimization problem while considering its relationship with a problem from Chap. 1. In Problem 1.2.2, setting $\mathbf{x} = (\mathbf{a}^\top, \mathbf{u}^\top)^\top \in \mathbb{R}^{2n}$ where $n \in \mathbb{N}$, we had the equality constraint $\mathbf{h}(\mathbf{x}) = \mathbf{K}(\mathbf{a})\mathbf{u} - \mathbf{p} = \mathbf{0}_{\mathbb{R}^n}$. If $2n$ is replaced with $d \in \mathbb{N}$, an optimization problem of the following type is obtained.

Problem 2.1.1 (Optimization Problem) Let $X = \mathbb{R}^d$ and assume that $f_0, f_1, \dots, f_m : X \rightarrow \mathbb{R}$ are given. Also assume that $\mathbf{h} = (h_1, \dots, h_n)^\top : X \rightarrow \mathbb{R}^n$ with respect to $n < d$ is given. Find \mathbf{x} satisfying

$$\min_{\mathbf{x} \in X} \{ f_0(\mathbf{x}) \mid f_1(\mathbf{x}) \leq 0, \dots, f_m(\mathbf{x}) \leq 0, \mathbf{h}(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^n} \}. \quad \square$$

Moreover, an optimization problem without equality constraint can be obtained from Problem 1.2.2 by regarding $\mathbf{a} \in X = \mathbb{R}^n$ as the design variable, and $\tilde{f}_i(\mathbf{a}) = f_i(\mathbf{a}, \mathbf{u}(\mathbf{a}))$. If $\mathbf{a} \in X = \mathbb{R}^n$ is rewritten as $\mathbf{x} \in X = \mathbb{R}^d$, and $\tilde{f}_i(\mathbf{a})$ is replaced by $f_i(\mathbf{x})$, then Problem 1.2.2 becomes an optimization problem as follows:

Problem 2.1.2 (Optimization Problem) Let $X = \mathbb{R}^d$ and suppose that $f_0, f_1, \dots, f_m : X \rightarrow \mathbb{R}$ are given. Find \mathbf{x} satisfying

$$\min_{\mathbf{x} \in X} \{ f_0(\mathbf{x}) \mid f_1(\mathbf{x}) \leq 0, \dots, f_m(\mathbf{x}) \leq 0 \}. \quad \square$$

Furthermore, let the set of admissible design variables (also referred to as the feasible set) be

$$S = \{ \mathbf{x} \in X \mid f_1(\mathbf{x}) \leq 0, \dots, f_m(\mathbf{x}) \leq 0 \}. \quad (2.1.1)$$

In this case, Problem 2.1.2 is equivalent to the following problem.

Problem 2.1.3 (Optimization Problem) Let $X = \mathbb{R}^d$ and suppose that $f_0, f_1, \dots, f_m : X \rightarrow \mathbb{R}$ are given. Let S be given by Eq. (2.1.1). Find \mathbf{x} satisfying

$$\min_{\mathbf{x} \in S} f_0(\mathbf{x}). \quad \square$$

In this chapter, let us look at conditions that are satisfied by minimizers of Problems 2.1.1, 2.1.2, and 2.1.3. Before beginning this topic, we would like to draw attention to a few definitions that will be used from here on. In this chapter, f_0 is called the objective function, \mathbf{h} is called an equality constraint function, and f_1, \dots and f_m denote inequality constraint functions. Two points are worth noting here.

The first is a cautionary remark regarding the direction of the inequality in the inequality constraint and its relationship to the optimization of the objective function. In particular, the maximization of f_0 is equivalent to the minimization of $-f_0$, and so we can so f_0 can be limited to minimization. Moreover, without loss of generality, f_1, \dots, f_m can be restricted to be non-positive.

The second remark regards the fact that for each $i \in \{1, \dots, n\}$ the equality constraint $h_i = 0$ is equivalent to imposing two simultaneous inequality constraints: $h_i \leq 0$ and $-h_i \leq 0$. Here, by replacing $h_i = 0$ in Problem 2.1.1 with $h_i \leq 0$ and $-h_i \leq 0$, Problem 2.1.1 can be reformulated as Problem 2.1.2. However in this chapter, we will take a detailed look at the relationship formed at the minimum of Problem 2.1.1 when the equality constraints are specified separately. The reason for this is as follows. In an optimal design problem, state equations always appear as equality constraints. Later on, these will become partial differential equations of boundary value type. In this case, the method for treating equality constraints shown in this chapter (the Lagrange multiplier or adjoint variable method) will be the guiding principle when considering equality constraints in the function space setting. Details related to this will be shown in Chap. 7.

Let us illustrate the optimization problems considered in this chapter in figures. Examples of minimizers in optimization problems when $X = \mathbb{R}^2$ are shown in Figs. 2.1, 2.2, 2.3. Here, g_0, g_1, g_2 and $\partial_X h_1$ denote the gradient (Definition 2.4.1) with respect to an arbitrary variation of f_0, f_1, f_2 , and h_1 at $x \in X$. The space to which these gradients belong is referred to as the dual space of X and is denoted by X' (Definition 4.4.5). However, since $X' = X$ in a finite-dimensional vector space, we can assume that $X' = \mathbb{R}^2$. When this figure is used in Chap. 7, X' is treated as a different vector space than X . Later on, whenever defining optimization problems,

Fig. 2.1 The minimum when all constraints are inactive

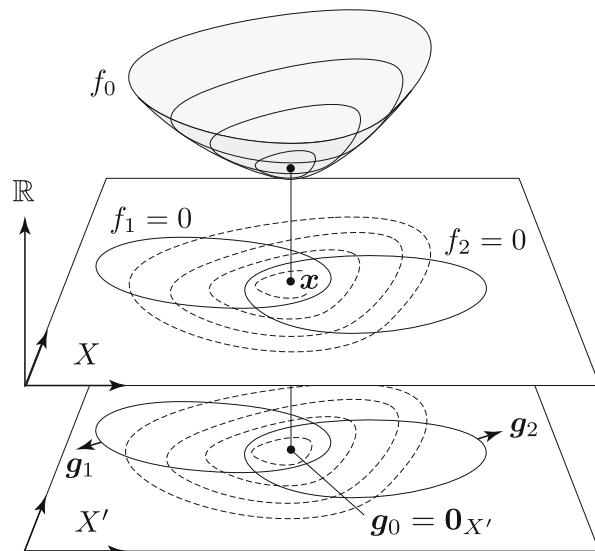


Fig. 2.2 The minimum when an equality constraint is active

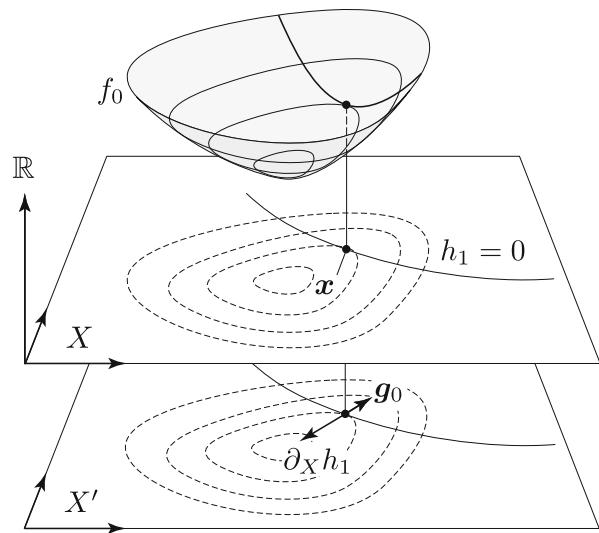
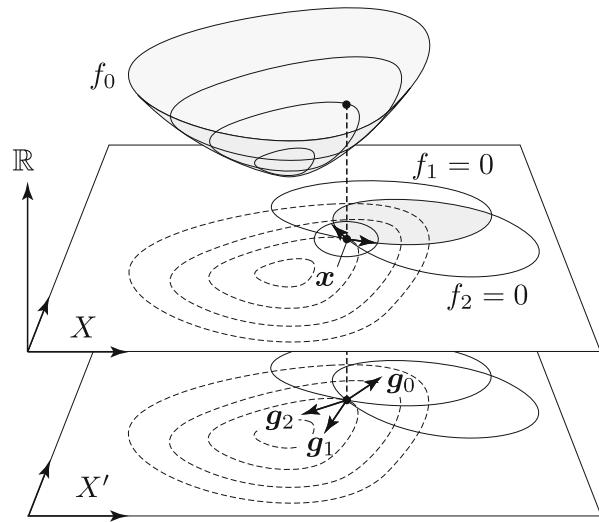


Fig. 2.3 The minimum when an inequality constraint is active

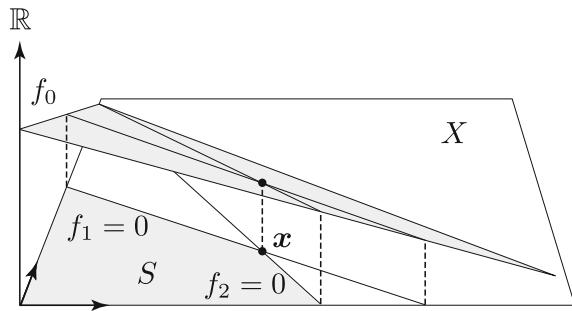


the variables and functions which appear should be referenced to these figures in order to understand the situation.

2.2 Classification of Optimization Problems

Next, in order to present a method for classifying optimization problems, let us focus on the properties of the functions used in Problems 2.1.1–2.1.3.

Fig. 2.4 The minimum point in a linear optimization problem



If f_0, \dots, f_m and h_1, \dots, h_n are all linear functions in Problem 2.1.1, or when f_0, \dots, f_m are all linear functions in Problems 2.1.2 and 2.1.3, then these problems are referred to as linear optimization problems or linear programming problems. Figure 2.4 shows the setting of a linear optimization problem when $X = \mathbb{R}^2$. Regarding the solution of linear optimization problems, methods such as the simplex method, which uses properties of linear functions, and the dual interior point method are well known. There are even cases when non-linear optimization problems can be solved after being changed into a linear optimization problem via successive linear approximation. However, the details will be omitted in this book since they will not be directly used.

On the other hand, in Problem 2.1.1, if some function within f_0, \dots, f_m or h_1, \dots, h_n is not linear, or if some function from f_0, \dots, f_m in Problem 2.1.2 or Problem 2.1.3 is not a linear, then these problems are referred to as non-linear optimization problems or as non-linear programming problems.

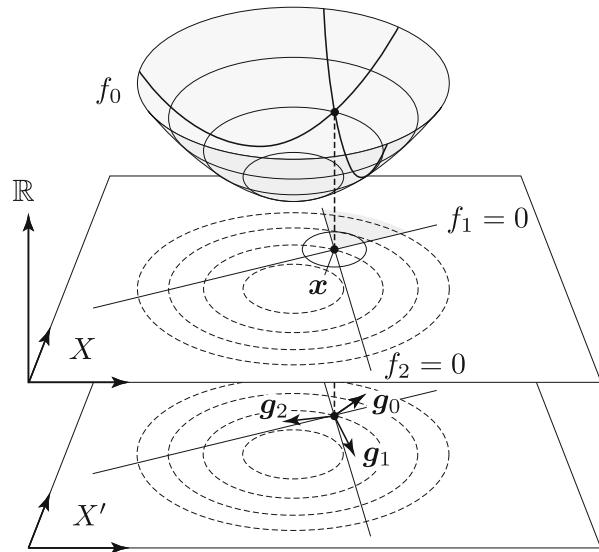
Moreover, when f_0 is a quadratic function and each of h_1, \dots, h_n and f_1, \dots, f_m are linear functions, or when f_0 in Problem 2.1.2 and Problem 2.1.3 is a quadratic function, and f_1, \dots, f_m are all linear functions, these problems are called quadratic optimization problems or quadratic programming problems. Figure 2.5 demonstrates a quadratic optimization problem when $X = X' = \mathbb{R}^2$.

Furthermore, when f_0, \dots, f_m are convex functions (Definition 2.4.3) and h_1, \dots, h_n are linear in Problem 2.1.1, or f_0, \dots, f_m are convex functions in Problem 2.1.2 or Problem 2.1.3, these problems are called convex optimization problems or convex programming problems.

Here we would like to consider a problem which can be a source of some confusion. As can be understood from Sect. 1.1.7, Problem 1.1.4 is a convex optimization problem. This is because when Problem 1.1.4 is rewritten in the form of Problem 2.1.2, the Hesse matrix H_0 of \tilde{f}_0 is positive definite (Theorem 2.4.6) and f_1 is a linear function with respect to the design variable (Theorem 2.4.4). However, Problem 1.1.4 does not look like a convex optimization problem when it is rewritten in the form of Problem 2.1.1.

This is because when we set $\mathbf{x} = (\mathbf{a}^\top, \mathbf{u}^\top)^\top$ and $\mathbf{h}(\mathbf{x}) = \mathbf{K}(\mathbf{a})\mathbf{u} - \mathbf{p} = \mathbf{0}_{\mathbb{R}^n}$ in Problem 2.1.1, $\mathbf{h}(\mathbf{x})$ is not a linear function of \mathbf{x} . Therefore it doesn't look like a convex optimization problem. As just described, optimization problems are

Fig. 2.5 Minimum point in a quadratic optimization problem



still convex when they are obtained by rewriting a problem which includes equality constraints into a form without the equality constraints.

Also, although the problem formulation is different from Problems 2.1.1–2.1.3, optimization problems equipped with multiple objective functions are called multi-objective optimization problems. For example, if S denotes a subset of $X = \mathbb{R}^d$ and $f_1, \dots, f_m : X \rightarrow \mathbb{R}$ are cost functions, then the problem becomes one in which we seek \mathbf{x} to satisfy

$$\min_{\mathbf{x} \in S} f_1(\mathbf{x}), \dots, \min_{\mathbf{x} \in S} f_m(\mathbf{x}).$$

We remark that minimizers need not exist for problems of this type and that, when the minimizer does not exist, the next best choice to use is the set of so-called Pareto solutions. Pareto solutions are defined as a set $P \subset S$ that satisfies the following conditions (see Fig. 2.6). Given an element \mathbf{y} of P , there does not exist an $\mathbf{x} \in S$ such that the following holds for all $i \in \{1, \dots, m\}$:

$$f_i(\mathbf{x}) \leq f_i(\mathbf{y}).$$

Moreover, for each fixed $i \in \{1, \dots, m\}$ there does not exist $\mathbf{x} \in S$ satisfying

$$f_i(\mathbf{x}) < f_i(\mathbf{y})$$

for all $\mathbf{y} \in P$. In order to select a point from a set of Pareto solutions, a selection criterion based on a concrete value system is needed. Nevertheless, discussions

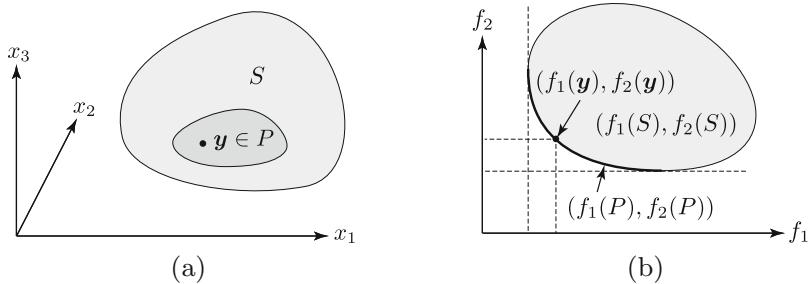


Fig. 2.6 A set of Pareto solutions, P . (a) Linear space of design variables. (b) Linear space of cost functions

surrounding such selection criteria will be avoided and multi-purpose optimization problems will not be considered in this book.

Furthermore, when the set of admissible design variables S is a set of discrete points (such as the integers) in Problem 2.1.3, this type of problem is referred to as a discrete optimization problem or a discrete programming problem. These problems display a property that is referred to as NP-hard, and exact solutions cannot be easily found. Moreover, special schemes are needed in order to obtain approximate solutions. These problems will also not be examined in this book.

2.3 Existence of a Minimum Point

Having defined optimization problems, let us now consider conditions under which minimizers exist in our problems. Although the following may seem obvious, failure to acknowledge such observations may lead to defining optimization problems for which no minimum is obtained.

Let us first note the difference between a local minimum point and a minimum point.

Definition 2.3.1 (Local and Global Minimizers) Let $X = \mathbb{R}^d$ and assume that S is non-empty subset of X . Also let $f : S \rightarrow \mathbb{R}$. When a neighborhood (a convex open set) B of $\mathbf{x} \in S$ exists and the following holds with respect to an arbitrary $\mathbf{y} \in B$:

$$f(\mathbf{x}) \leq f(\mathbf{y}),$$

then we say that $f(\mathbf{x})$ obtains a local minimum value at \mathbf{x} , and that \mathbf{x} is a local minimizer. When the above inequality holds with respect to an arbitrary $\mathbf{y} \in S$, then we say that $f(\mathbf{x})$ obtains its minimum value at \mathbf{x} , and that \mathbf{x} is the global minimizer.

□

Fig. 2.7 A local minimizer x and the global minimizer y of function f

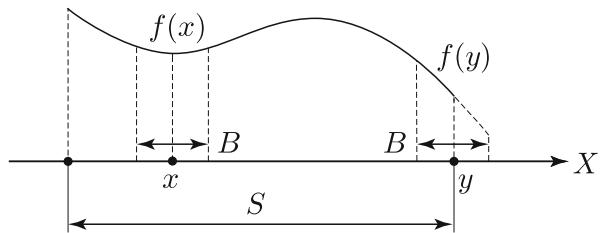


Figure 2.7 shows a local minimum x and the global minimum y of a function f when $X = \mathbb{R}$ and $S \subset X$ is a bounded closed set. Definitions of terminology such as neighborhood and open set can be found in Sect. A.1.1. When defining global minimizers as in Definition 2.3.1, the well-known Weierstrass's theorem gives sufficient conditions for their existence (cf. [161, Theorem 13, p. 27], [21, Section 22.6, p. 154], or [139, Theorem 4.16, p. 89]).

Theorem 2.3.2 (Weierstrass's Theorem) *Let S be a bounded closed subset of $X = \mathbb{R}^d$, and $f_0 : S \rightarrow \mathbb{R}$ be a continuous function in Problem 2.1.3. Then there exists a global minimizer of f_0 in S .* \square

Note that continuous functions are defined in Sect. A.1.2, and that Theorem 2.3.2 still holds when continuity is replaced by lower semi-continuity (Sect. A.1.2). However, given that lower semi-continuity will not be required in our future developments, this extension is omitted.

Let us take a look at cases where global minimizers fail to exist when S is not a bounded closed set in Problem 2.1.3. In Problem 1.1.4, the following assumption was made:

$$S = \left\{ \mathbf{a} \in X = \mathbb{R}^2 \mid \mathbf{a} \geq \mathbf{a}_0, f_1(\mathbf{a}) = l(a_1 + a_2) - c_1 \leq 0 \right\}.$$

This set is bounded and closed. However, if this is replaced by

$$S = \left\{ \mathbf{a} \in X = \mathbb{R}^2 \mid \mathbf{a} > \mathbf{0}_{\mathbb{R}^2}, f_1(\mathbf{a}) = l(a_1 + a_2) - c_1 \leq 0 \right\}$$

then it is no longer a bounded closed set. For example, when $p_1 \neq 0$ and $p_2 = 0$, it follows that $a_1 = c_1/2l$ and $a_2 = 0$ express the minimum of f_0 . However, \mathbf{a} is not included in S hence there are no minimizers within S .

Moreover, when the underlying function is discontinuous, or doesn't uniquely determine its values, there is no guarantee that a minimum value will exist. None of the functions looked at so far has these qualities. However, if care is not taken, functional optimization problems may well result in having to deal with functions whose values are not uniquely defined. In Chap. 4, the linear space in which the design variables are defined is continuous (complete), and conditions under which the cost function are continuous will be considered.

2.4 Differentiation and Convex Functions

Before addressing the optimization problem, let us look at the basic methods of differentiation used in optimization theory, including the definition of a convex functions. We will use the following definition for the derivative of a function f .

Definition 2.4.1 (Gradient and Hesse Matrix) Let $X = \mathbb{R}^d$ and suppose a function $f : B \rightarrow \mathbb{R}$ is defined in the neighborhood $B \subset X$ of $\mathbf{x} \in X$. When $\mathbf{y} = (y_1, \dots, y_d)^\top \in X$ is arbitrary and

$$\begin{aligned}\partial_X f(\mathbf{x}) &= \begin{pmatrix} \lim_{y_1 \rightarrow 0} (f(\mathbf{x} + (y_1, 0, \dots, 0)^\top) - f(\mathbf{x})) / y_1 \\ \vdots \\ \lim_{y_d \rightarrow 0} (f(\mathbf{x} + (0, \dots, y_d)^\top) - f(\mathbf{x})) / y_d \end{pmatrix} \\ &= \begin{pmatrix} \partial f / \partial x_1 \\ \vdots \\ \partial f / \partial x_d \end{pmatrix}(\mathbf{x}) = \mathbf{g}(\mathbf{x})\end{aligned}$$

is an element of $X' = \mathbb{R}^d$, then f is said to be differentiable at \mathbf{x} , and

$$f'(\mathbf{x})[\mathbf{y}] = \mathbf{g}(\mathbf{x}) \cdot \mathbf{y}$$

is called the derivative, or the total derivative, of f at \mathbf{x} and $\mathbf{g}(\mathbf{x})$ is the gradient of f at \mathbf{x} . Likewise, when we can resolve

$$\partial_X \partial_X^\top f(\mathbf{x}) = \begin{pmatrix} \partial^2 f / (\partial x_1 \partial x_1) & \dots & \partial^2 f / (\partial x_1 \partial x_d) \\ \vdots & \ddots & \vdots \\ \partial^2 f / (\partial x_d \partial x_1) & \dots & \partial^2 f / (\partial x_d \partial x_d) \end{pmatrix}(\mathbf{x}) = \mathbf{H}(\mathbf{x})$$

as an element of $\mathbb{R}^{d \times d}$, then f is said to be second-order differentiable at \mathbf{x} . Moreover,

$$f''(\mathbf{x})[\mathbf{y}_1, \mathbf{y}_2] = \mathbf{y}_2 \cdot (\mathbf{H}(\mathbf{x}) \mathbf{y}_1)$$

is referred to as the second-order derivative of f at \mathbf{x} with respect to arbitrary variations $\mathbf{y}_1, \mathbf{y}_2 \in X$ from \mathbf{x} , and $\mathbf{H}(\mathbf{x})$ is referred to as the Hesse matrix. \square

In this book, the set of functions $f : X \rightarrow \mathbb{R}$ whose first $k \in \{0, 1, 2, \dots\}$ derivatives are continuous over X will be denoted by $C^k(X; \mathbb{R})$ (Definition 4.2.2). Moreover, for simplicity of notation, $\partial_X f, \partial_X f_0, \dots, \partial_X f_m$ will be denoted by $\mathbf{g}, \mathbf{g}_0, \dots, \mathbf{g}_m$ respectively, and $\partial_X \partial_X^\top f, \partial_X \partial_X^\top f_0, \dots, \partial_X \partial_X^\top f_m$ will be denoted by $\mathbf{H}, \mathbf{H}_0, \dots, \mathbf{H}_m$ respectively. Also, when \mathbf{x} is an element of X , we note that $\partial_X f$ will also be written as f_x . When considering partial differential equations, etc.,

we remark that $\partial_X f$ will be expressed by ∇f . Various types of derivatives will be defined going forward, so methods for their expression will be needed. Definitions will be given in each situation to avoid confusion.

2.4.1 Taylor's Theorem

Used in all kinds of situations hereafter, Taylor's theorem is shown below.

Theorem 2.4.2 (Taylor's Theorem) *Let $X = \mathbb{R}^d$. Suppose that a function $f \in C^2(B; \mathbb{R})$ is defined in a neighborhood B of $\mathbf{a} \in X$. If $\mathbf{y} = \mathbf{b} - \mathbf{a}$ with respect to an arbitrary $\mathbf{b} \in B$, then there exists $\theta \in (0, 1)$ satisfying*

$$f(\mathbf{b}) = f(\mathbf{a}) + \mathbf{g}(\mathbf{a}) \cdot \mathbf{y} + \frac{1}{2!} \mathbf{y} \cdot (\mathbf{H}(\mathbf{a} + \theta \mathbf{y}) \mathbf{y}). \quad (2.4.1)$$

□

Proof First assume that $X = \mathbb{R}$. Let \mathbf{a} , \mathbf{b} and \mathbf{y} in this case be denoted by a , b and y respectively. Given $x \in B$, let

$$h(x) = f(b) - \left\{ f(x) + f'(x)(b-x) + k(b-x)^2 \right\},$$

where we have written df/dx as f' . Moreover, the constant k is determined such that

$$h(a) = h(b) = 0.$$

We obtain

$$\begin{aligned} h'(x) &= -f'(x) - f''(x)(b-x) + f'(x) + 2k(b-x) \\ &= -f''(x)(b-x) + 2k(b-x). \end{aligned}$$

By Rolle's theorem (the mean value theorem when $h(a) = h(b)$), there exists c in (a, b) satisfying

$$h'(c) = 0.$$

Hence, we can write $c = a + \theta y$ and obtain

$$k = \frac{1}{2} f''(a + \theta y).$$

Substituting this result into $h(a) = 0$ yields the result of the theorem.

Next let $X = \mathbb{R}^2$. Consider the following function of $t \in \mathbb{R}$ with respect to an arbitrary $\mathbf{a} = (a_1, a_2)^\top$ and $\mathbf{y} = (y_1, y_2)^\top$:

$$\phi(t) = f(\mathbf{a} + t\mathbf{y})$$

$$= f(\mathbf{a}) + t \left(y_1 \frac{\partial}{\partial x_1} + y_2 \frac{\partial}{\partial x_2} \right) f(\mathbf{a}) + \frac{t^2}{2} \left(y_1 \frac{\partial}{\partial x_1} + y_2 \frac{\partial}{\partial x_2} \right)^2 f(\mathbf{a} + \theta\mathbf{y}).$$

By Taylor expanding $\phi(t)$ around $t = 0$ (as a function of one variable, $X = \mathbb{R}$) the value of $\phi(1)$ can be written as

$$\begin{aligned} \phi(1) &= f(\mathbf{b}) \\ &= f(\mathbf{a}) + \left(y_1 \frac{\partial}{\partial x_1} + y_2 \frac{\partial}{\partial x_2} \right) f(\mathbf{a}) \\ &\quad + \frac{1}{2} \left(y_1 \frac{\partial}{\partial x_1} + y_2 \frac{\partial}{\partial x_2} \right)^2 f(\mathbf{a} + \theta\mathbf{y}) \\ &= f(\mathbf{a}) + \left\{ \left(\frac{\partial}{\partial x_1} \frac{\partial}{\partial x_2} \right) f(\mathbf{a}) \right\} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \\ &\quad + \frac{1}{2} (y_1 \ y_2) \left\{ \left(\frac{\partial}{\partial x_1} \frac{\partial}{\partial x_2} \right) \left(\frac{\partial}{\partial x_1} \frac{\partial}{\partial x_2} \right) f(\mathbf{a} + \theta\mathbf{y}) \right\} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}. \end{aligned} \quad \square$$

When $X = \mathbb{R}^d$, the above equation becomes Eq. (2.4.1).

Taylor's theorem can also be expressed with respect to arbitrary derivative orders. However, our expression stops at the second-order derivative of the function because the notation for higher-order differentials defined in \mathbb{R}^d have not been defined. This notation will be presented in Definition 4.2.2. Moreover, we also refer to terms such as $\mathbf{y} \cdot (\mathbf{H}(\mathbf{a} + \theta\mathbf{y}) \mathbf{y}) / 2!$ containing θ as remainder terms. Also, we will make use of the Bachmann–Landau small- o symbol in this book. This allows us to write Eq. (2.4.1) also as

$$f(\mathbf{a} + \mathbf{y}) = f(\mathbf{a}) + \mathbf{g}(\mathbf{a}) \cdot \mathbf{y} + \frac{1}{2!} \mathbf{y} \cdot (\mathbf{H}(\mathbf{a}) \mathbf{y}) + o\left(\|\mathbf{y}\|_{\mathbb{R}^d}^2\right), \quad (2.4.2)$$

where $o(\epsilon)$ is defined to be a function such that $\lim_{\epsilon \rightarrow 0} o(\epsilon) / \epsilon = 0$. This equation is referred to as a Taylor expansion of f around \mathbf{a} .

2.4.2 Convex Functions

Next let us take a look at a few basic definitions and results regarding convex functions. As we will show in Theorem 2.5.6, local and global minimizers coincide in the case of convex optimization problems. For this reason, convexity of functions is an important and useful property in optimization theory. The definition of a convex function is as follows.

Definition 2.4.3 (Convex Functions) Let $X = \mathbb{R}^d$ and S be a non-empty subset of X . A function $f : S \rightarrow \mathbb{R}$ is said to be convex if the following holds for arbitrary $\mathbf{x}, \mathbf{y} \in S$ and $\theta \in (0, 1)$:

$$(1 - \theta)\mathbf{x} + \theta\mathbf{y} \in S, \quad (2.4.3)$$

$$f((1 - \theta)\mathbf{x} + \theta\mathbf{y}) \leq (1 - \theta)f(\mathbf{x}) + \theta f(\mathbf{y}). \quad (2.4.4)$$

When the direction of the inequality is reversed, f is called a concave function. \square

Equations (2.4.3) and (2.4.4) are conditions which signify the convexity of a set and the convexity of a function, respectively. Figures 2.8 and 2.9 illustrate the case of convex and non-convex S . Moreover, if S is a set of integer-valued vectors (such as in Fig. 2.9c), the optimization problem becomes a discrete programming problem. A difficulty regarding these problems can be thought to lie in the fact that, when the admissible set consists of such vectors, points at which the gradient of the cost function are $\mathbf{0}_X$ need not be included in the admissible set. Figure 2.10 shows

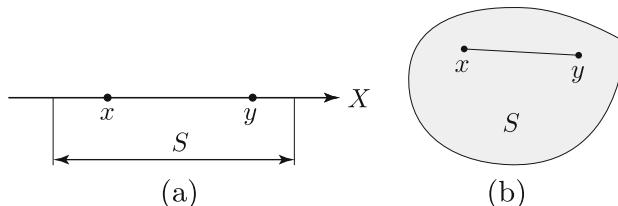


Fig. 2.8 S as a convex subset of the linear space X . (a) $X = \mathbb{R}$. (b) $X = \mathbb{R}^2$

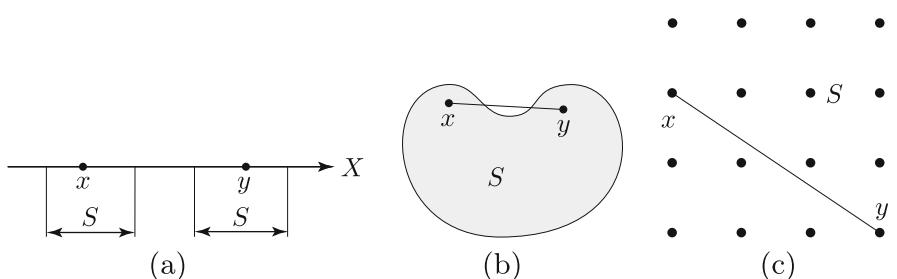


Fig. 2.9 S as a non-convex subset of the linear space X . (a) $X = \mathbb{R}$. (b) $X = \mathbb{R}^2$. (c) $X = \mathbb{R}^2$

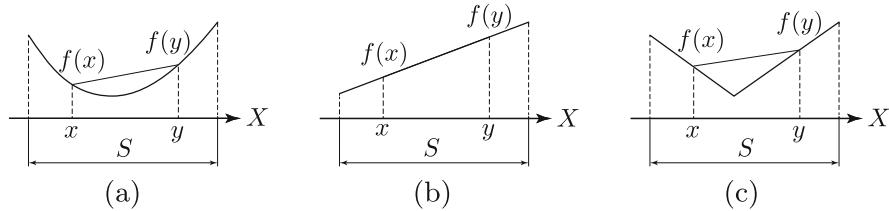


Fig. 2.10 Examples of convex functions ($X = \mathbb{R}$). (a) $f \in C^1(S; \mathbb{R})$. (b) f : linear function. (c) $f \in C^0(S; \mathbb{R})$

an example when f is a convex function. We also note that, even if the derivative is non-continuous, convexity can still hold.

If a convex function is first-order differentiable, the following results can be obtained.

Theorem 2.4.4 (Convexity and First-Order Differentiation) *Suppose $f \in C^1(S; \mathbb{R})$ and let $S \subseteq X$ be an open convex set of $X = \mathbb{R}^d$. A necessary and sufficient condition for f to be a convex function is for the following inequality to hold for arbitrary $\mathbf{x}, \mathbf{y} \in S$:*

$$\mathbf{g}(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}) \leq f(\mathbf{y}) - f(\mathbf{x}). \quad (2.4.5)$$

□

Proof We will first show the necessity (that Eq. (2.4.5) holds when f is a convex function). Since f is a convex function, one has

$$f((1 - \theta)\mathbf{x} + \theta\mathbf{y}) = f(\mathbf{x} + \theta(\mathbf{y} - \mathbf{x})) \leq f(\mathbf{x}) + \theta(f(\mathbf{y}) - f(\mathbf{x})),$$

where $\mathbf{x}, \mathbf{y} \in S$ and $\theta \in (0, 1)$ are arbitrary. Hence, it follows that

$$\frac{f(\mathbf{x} + \theta(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{\theta} \leq f(\mathbf{y}) - f(\mathbf{x}). \quad (2.4.6)$$

When $\theta \rightarrow 0$, we obtain Eq. (2.4.5).

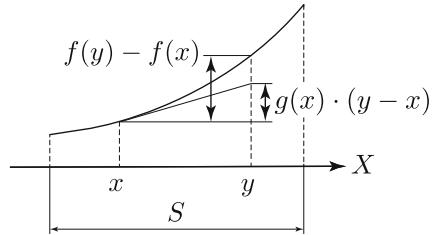
Next we will show the sufficiency (that is, if Eq. (2.4.5) holds, then f is a convex function). If we let $\mathbf{x} = (1 - \theta)\mathbf{z} + \theta\mathbf{w}$ and $\mathbf{y} = \mathbf{z}$, then Eq. (2.4.5) becomes

$$f(\mathbf{z}) - f(\mathbf{x}) \geq \mathbf{g}(\mathbf{x}) \cdot (\mathbf{z} - \mathbf{x}). \quad (2.4.7)$$

Similarly, if $\mathbf{x} = (1 - \theta)\mathbf{z} + \theta\mathbf{w}$ and $\mathbf{y} = \mathbf{w}$, Eq. (2.4.5) becomes

$$f(\mathbf{w}) - f(\mathbf{x}) \geq \mathbf{g}(\mathbf{x}) \cdot (\mathbf{w} - \mathbf{x}). \quad (2.4.8)$$

Fig. 2.11 Convexity and the first-order derivative ($X = \mathbb{R}$)



Multiplying Eq. (2.4.7) by $(1 - \theta)$, and Eq. (2.4.8) by θ and taking their sum yields:

$$(1 - \theta) f(z) + \theta f(w) - f(x) \geq g(x) \cdot \{(1 - \theta)z + \theta w - x\} = 0.$$

In other words, since S is a convex set, we obtain the convexity of f :

$$(1 - \theta) f(z) + \theta f(w) \geq f((1 - \theta)z + \theta w).$$

□

Figure 2.11 illustrates the content of Theorem 2.4.4.

Additionally, properties of Hesse matrices can be obtained when convex functions are second-order differentiable. In order to derive the result, let us define the notion of a positive definite real symmetric matrix.

Definition 2.4.5 (Positive Definite Real Symmetric Matrix) Let $A = A^\top \in \mathbb{R}^{d \times d}$. Then A is said to be positive definite if there exists $\alpha > 0$ satisfying

$$\mathbf{x} \cdot (A\mathbf{x}) \geq \alpha \|\mathbf{x}\|_{\mathbb{R}^d}^2,$$

for all $\mathbf{x} \in \mathbb{R}^d$. When there only exists $\alpha \geq 0$ satisfying the above, then A is said to be semi-positive definite. Similarly, when $\alpha > 0$ exists and

$$\mathbf{x} \cdot (A\mathbf{x}) \leq -\alpha \|\mathbf{x}\|_{\mathbb{R}^d}^2,$$

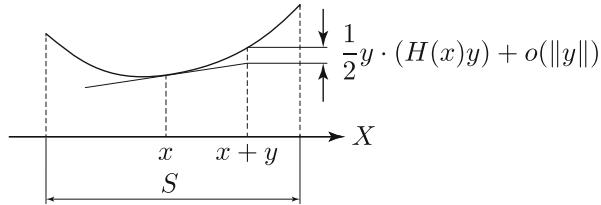
for all $\mathbf{x} \in \mathbb{R}^d$, then A is said to be negative definite. In the case that there only exists $\alpha \geq 0$ satisfying the above, we say that A is semi-negative definite. □

If A in Definition 2.4.5 is positive definite, then its eigenvalues are all positive and α is equal to their minimum value. Moreover, if A is negative definite, then all its eigenvalues are negative and $-\alpha$ is equal to their maximum value. The reader is encouraged to confirm these facts in Exercise 2.1.

When a convex function is second-order differentiable, these definitions can be used to formulate the following result (illustrated in Fig. 2.12).

Theorem 2.4.6 (Convexity and Second-Order Differentiation) Let $X = \mathbb{R}^d$, $S \subseteq X$ be an open convex set and $f \in C^2(S; \mathbb{R})$. Then the necessary and sufficient condition for f to be a convex function is that the Hesse matrix $H(\mathbf{x})$ is semi-positive definite with respect to arbitrary $\mathbf{x} \in S$. □

Fig. 2.12 Convexity and the second-order derivative ($X = \mathbb{R}$)



Proof We will first show necessity. Since f is a convex function, Eq. (2.4.6) holds for arbitrary $\mathbf{x}, \mathbf{y} \in S$ and $\theta \in (0, 1)$. Since $f \in C^2(S; \mathbb{R})$, given $\mathbf{x}, \mathbf{y} \in S$ there exists $\vartheta \in (0, 1)$ such that the right-hand-side of Eq. (2.4.6) can be written as

$$f(\mathbf{y}) - f(\mathbf{x}) = \mathbf{g}(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}) + \frac{1}{2} (\mathbf{y} - \mathbf{x}) \cdot \{\mathbf{H}((1 - \vartheta)\mathbf{x} + \vartheta\mathbf{y})(\mathbf{y} - \mathbf{x})\}. \quad (2.4.9)$$

Writing $\theta = \vartheta$, the left-hand side of Eq. (2.4.6) can be expressed as

$$\begin{aligned} & \frac{f(\mathbf{x} + \vartheta(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{\vartheta} \\ &= \mathbf{g}(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}) + \frac{1}{2} \vartheta(\mathbf{y} - \mathbf{x}) \cdot \{\mathbf{H}((1 - \vartheta)\mathbf{x} + \vartheta\mathbf{y})(\mathbf{y} - \mathbf{x})\}. \end{aligned} \quad (2.4.10)$$

Therefore, substituting Eqs. (2.4.9) and (2.4.10) into Eq. (2.4.6) when $\theta = \vartheta$ we obtain:

$$(1 - \vartheta)(\mathbf{y} - \mathbf{x}) \cdot \{\mathbf{H}((1 - \vartheta)\mathbf{x} + \vartheta\mathbf{y})(\mathbf{y} - \mathbf{x})\} \geq 0.$$

When $\vartheta \rightarrow 0$, $\mathbf{H}(\mathbf{x})$ is semi-positive definite.

Next let us show sufficiency. Given $\mathbf{x}, \mathbf{y} \in S$ there exists $\vartheta \in (0, 1)$ such that Eq. (2.4.9) is satisfied. Since $\mathbf{H}((1 - \vartheta)\mathbf{x} + \vartheta\mathbf{y})$ is semi-positive definite, the second term on the right-hand side of Eq. (2.4.9) is non-zero, and thus

$$f(\mathbf{y}) - f(\mathbf{x}) \geq \mathbf{g}(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}).$$

From Theorem 2.4.4 it follows that f is a convex function. \square

2.4.3 Exercises in Differentiation and Convex Functions

Let us make use of the previous theorems involving differentiation and convex functions in relation to simple problems from mechanics. Let us first confirm that the potential energy in the one-degree-of-freedom spring system considered in Exercise 1.1.1 is positive definite.

Exercise 2.4.7 (Potential Energy of a 1DOF Spring System) The potential energy in a one-degree-of-freedom spring system (such as is shown in Fig. 1.2) is given by

$$\pi(u) = \int_0^u (kv - p) \, dv = \frac{1}{2}ku^2 - pu.$$

Show that π is a convex function. \square

Answer With respect to π , we obtain

$$\frac{d^2\pi}{du^2} = k > 0.$$

By Theorem 2.4.6, π is a convex function. \square

Let us also confirm that the potential energy of a two-degree-of-freedom spring system is a convex function.

Exercise 2.4.8 (Potential Energy of a 2DOF Spring System) The potential energy of a two-degree-of-freedom spring system (such as is shown in Fig. 1.3) is given by

$$\pi(u) = \frac{1}{2}k_1u_1^2 + \frac{1}{2}k_2(u_2 - u_1)^2 - (p_1u_1 + p_2u_2).$$

Show that π is a convex function. \square

Answer The Hesse matrix of π is

$$\mathbf{H} = \begin{pmatrix} k_1 + k_2 & -k_2 \\ -k_2 & k_2 \end{pmatrix}.$$

The eigenvalues of \mathbf{H} are λ satisfying

$$\det \begin{vmatrix} k_1 + k_2 - \lambda & -k_2 \\ -k_2 & k_2 - \lambda \end{vmatrix} = 0,$$

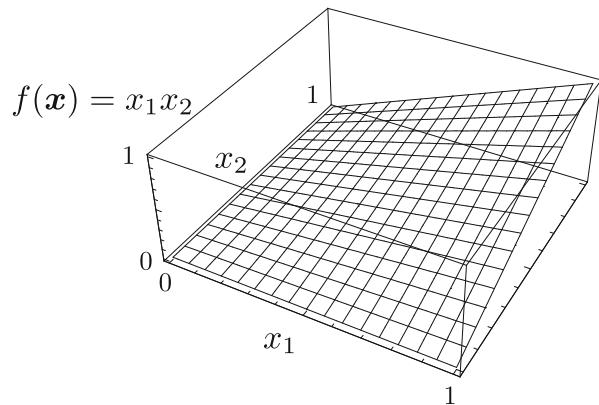
which leads to

$$\lambda_1, \lambda_2 = \frac{k_1 + 2k_2 \pm \sqrt{(k_1 + 2k_2)^2 - 4k_1k_2}}{2}.$$

It follows that λ_1 and λ_2 are greater than zero whenever $k_1, k_2 > 0$. Therefore, since all of its eigenvalues are positive, \mathbf{H} is positive definite (Theorem A.2.1). By Theorem 2.4.6, π is a convex function. \square

Fig. 2.13 The function

$$f(\mathbf{x}) = x_1 x_2$$



Let us finish this section by taking a look at an example involving a familiar function which is not convex.

Exercise 2.4.9 (Area of a Rectangle) Let $x_1 \in \mathbb{R}$ and $x_2 \in \mathbb{R}$ denote the length and width of a rectangle. Show that the area $f(\mathbf{x}) = x_1 x_2$ is not a convex function. \square

Answer Upon substituting $\mathbf{x} = (1, 0)^\top$ and $\mathbf{y} = (0, 1)^\top$ into Eq. (2.4.4), we obtain

$$\begin{aligned} f((1-\theta)\mathbf{x} + \theta\mathbf{y}) &= \{(1-\theta)x_1 + \theta y_1\} \{(1-\theta)x_2 + \theta y_2\} = (1-\theta)\theta \\ &\geq (1-\theta)f(\mathbf{x}) + \theta f(\mathbf{y}) = 0. \end{aligned}$$

Therefore, f is not a convex function (see Fig. 2.13). \square

2.5 Unconstrained Optimization Problems

We will now consider various cases of Problems 2.1.1–2.1.3 and look at foundational theorems of optimization theory. We will first consider the case when there are no constraints. This can be thought of as the case where all of the constraints are inactive. Namely, in Problem 2.1.3, that the minimum point is an interior point of $S \subseteq X = \mathbb{R}^d$. In this section, we will assume that S is an open set and denote f_0 by f in order to consider the following problem.

Problem 2.5.1 (Unconstrained Optimization Problems) Let $X = \mathbb{R}^d$ and S be an open subset of X . When $f : S \rightarrow \mathbb{R}$ is given, find \mathbf{x} satisfying

$$\min_{\mathbf{x} \in S} f(\mathbf{x}).$$

\square

2.5.1 A Necessary Condition for Local Minimizers

The conditions satisfied when $\mathbf{x} \in S$ is a local minimizer are referred to as necessary conditions for local minimization. When f is first-order differentiable, its derivative is defined and we obtain the following result expressing necessary conditions satisfied by local minimizers.

Theorem 2.5.2 (Necessary Conditions for Local Minimizers) *Let $f \in C^1(S; \mathbb{R})$ in Problem 2.5.1. If $\mathbf{x} \in S$ is a local minimizer, then*

$$\mathbf{g}(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^d}. \quad (2.5.1)$$

Equivalently, for all $\mathbf{y} \in X$

$$\mathbf{g}(\mathbf{x}) \cdot \mathbf{y} = 0. \quad (2.5.2)$$

□

Proof Suppose for contradiction that $\mathbf{g}(\mathbf{x}) \neq \mathbf{0}_{\mathbb{R}^d}$. If we set $\mathbf{y} = -\mathbf{g}(\mathbf{x})$, then $\mathbf{g}(\mathbf{x}) \cdot \mathbf{y} = -\|\mathbf{g}(\mathbf{x})\|_{\mathbb{R}^d}^2 < 0$. Since \mathbf{g} is continuous, there exists t_1 such that

$$\mathbf{g}(\mathbf{x} + t\mathbf{y}) \cdot \mathbf{y} < 0$$

for all $t \in [0, t_1]$. By the mean value theorem, given $t \in (0, t_1]$ there exists $\theta \in (0, 1)$ such that

$$f(\mathbf{x} + t\mathbf{y}) = f(\mathbf{x}) + t\mathbf{g}(\mathbf{x} + \theta t\mathbf{y}) \cdot \mathbf{y}.$$

Since $\theta t \in (0, t_1)$ we have $\mathbf{g}(\mathbf{x} + \theta t\mathbf{y}) \cdot \mathbf{y} < 0$. Substituting this relation into the above equation yields a contradiction $f(\mathbf{x} + t\mathbf{y}) < f(\mathbf{x})$. □

Theorem 2.5.2, signifies that vector equalities (such as Eq. (2.5.1)) are equivalent to the condition that the inner product with any arbitrary vector (such as in Eq. (2.5.2)) is 0. Equivalence is obtained because Eq. (2.5.2) is required to hold for all $\mathbf{y} \in X$. Such expressions involving arbitrary vectors will appear frequently in this book, and it is thus important for the reader to understand this notion here.

If f is second-order differentiable, its Hesse matrix can be defined and the following results can be obtained.

Theorem 2.5.3 (2nd-Order Necessary Condition for a Local Minimizer) *Consider Problem 2.5.1 and let $f \in C^2(S; \mathbb{R})$. If $\mathbf{x} \in S$ is a local minimizer, then the Hesse matrix $\mathbf{H}(\mathbf{x})$ is semi-positive definite.* □

Proof If \mathbf{x} is a local minimizer, then by Definition 2.3.1 there exists a neighborhood $B \subset S$ of \mathbf{x} such that the following holds for all $\mathbf{y} \in B$:

$$f(\mathbf{y}) - f(\mathbf{x}) \geq 0.$$

Additionally, since \mathbf{x} is a local minimum, it follows from Taylor's theorem that

$$f(\mathbf{y}) - f(\mathbf{x}) = \frac{1}{2}(\mathbf{y} - \mathbf{x}) \cdot \{\mathbf{H}(\mathbf{x})(\mathbf{y} - \mathbf{x})\} + o\left(\|\mathbf{y} - \mathbf{x}\|_{\mathbb{R}^d}^2\right).$$

If $\mathbf{z} \in X$ is arbitrary and we multiply both sides of the above by $2\mathbf{z}/\|\mathbf{y} - \mathbf{x}\|_{\mathbb{R}^d}^2$ and take $\mathbf{y} \rightarrow \mathbf{x}$, then we obtain

$$\mathbf{z} \cdot (\mathbf{H}(\mathbf{x})\mathbf{z}) \geq 0.$$

□

2.5.2 Sufficient Conditions for Local Minimizers

Next, let us take a look at conditions guaranteeing when $\mathbf{x} \in S$ is a local minimum. In order to do so, we now give the definition of a stationary point.

Definition 2.5.4 (Stationary Point) Let $S \subseteq X$ be an open set and \mathbf{x} an element of S . When

$$\mathbf{g}(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^d}$$

we say that \mathbf{x} is a stationary point.

□

If f is second-order differentiable, then the following sufficient condition for attaining a local minimum can be obtained.

Theorem 2.5.5 (2nd-Order Sufficient Condition for a Local Minimizer) Consider Problem 2.5.1 and let $f \in C^2(S; \mathbb{R})$. If $\mathbf{x} \in S$ is a stationary point and the Hesse matrix $\mathbf{H}(\mathbf{x})$ is positive definite, then \mathbf{x} is a local minimizer. □

Proof Let $B \subset S$ be a neighborhood around a stationary point \mathbf{x} . Given $\mathbf{x} + \mathbf{y} \in B$ there exists $\theta \in (0, 1)$ such that

$$f(\mathbf{x} + \mathbf{y}) - f(\mathbf{x}) = \frac{1}{2}\mathbf{y} \cdot (\mathbf{H}(\mathbf{x} + \theta\mathbf{y})\mathbf{y}).$$

Since $\mathbf{H}(\mathbf{x} + \theta\mathbf{y})$ is positive definite, the result follows:

$$f(\mathbf{x} + \mathbf{y}) > f(\mathbf{x}).$$

□

2.5.3 Sufficient Conditions for Global Minimizers

The previous section established conditions satisfied by local minimizers. Let us now take a look at conditions established in the case of global minimizers. In particular, if Problem 2.5.1 is a convex optimization problem, then the following result can be obtained.

Theorem 2.5.6 (Sufficient Conditions for a Global Minimizer) *In Problem 2.5.1, let $S \subseteq X$ be a non-empty open convex set and $f : S \rightarrow \mathbb{R}$ be a convex function. If $\mathbf{x} \in S$ is a local minimizer, then \mathbf{x} yields the minimum over S . \square*

Proof If \mathbf{x} is a local minimizer, then there exists a neighborhood $B \subset S$ around \mathbf{x} such that \mathbf{x} yields the minimum value over B . If we suppose that there exists another minimizer $\mathbf{y} \in S$ which is different from \mathbf{x} , then for sufficiently small $\theta \in (0, 1)$ we can find $\mathbf{z} \in B$ such that

$$(1 - \theta) f(\mathbf{x}) + \theta f(\mathbf{y}) \geq f((1 - \theta) \mathbf{x} + \theta \mathbf{y}) = f(\mathbf{z}).$$

This is contrary to the fact that \mathbf{x} is a local minimizer. Hence, \mathbf{x} is the only local minimizer. \square

2.5.4 Example of Unconstrained Optimization Problem

As in the previous section, let us now confirm the results thus far in relation to unconstrained problems involving systems of springs. Let us first consider a one-degree-of-freedom spring system.

Exercise 2.5.7 (Force Equilibrium Equation in a 1DOF Spring System) Show that if u satisfies the force equilibrium equation of the one-degree-of-freedom spring system shown in Fig. 1.2, then it minimizes the potential energy π in Exercise 2.4.7. \square

Answer By Exercise 2.4.7, π is a convex function. If u satisfies the force equilibrium equation, then by Theorem 2.5.5 it is a local minimizer. By Theorem 2.5.6, it is also a global minimizer. \square

Next, let us treat a system involving multiple degrees-of-freedom.

Exercise 2.5.8 (Force Equilibrium Equation in a 2DOF Spring System) Show that if \mathbf{u} satisfies the force equilibrium equation of the two-degree-of-freedom spring system shown in Fig. 1.3, then it minimizes the potential energy π in Exercise 2.4.8. \square

Answer By Exercise 2.4.8, π is a convex function. If \mathbf{u} satisfies the force equilibrium equation, then by Theorem 2.5.5 it is a local minimizer. By Theorem 2.5.6, it is also a global minimizer. \square

2.5.5 Considerations Relating to the Solutions of Unconstrained Optimization Problems

Combining the results from this section with results which will be shown later allows us to conclude the following about solutions of the unconstrained optimization problem (Problem 2.5.1).

- (1) By Theorem 2.5.2, if a point is a local minimizer, it is also a stationary point. Hence stationary points are candidates for local minimizers.
- (2) If after obtaining a stationary point \mathbf{x} the Hesse matrix $\mathbf{H}(\mathbf{x})$ is found to be positive definite, then by Theorem 2.5.5 \mathbf{x} can be deemed to be a local minimizer.
- (3) If f is a convex function, any stationary point \mathbf{x} which is also a local minimizer is necessarily a global minimizer by Theorem 2.5.6.
- (4) When the convexity of f is unknown, local minimizers can be sought using various trial points and optimization methods developed in Chap. 3, amongst which the global minimizer can be found.

2.6 Optimization Problems with Equality Constraints

We will now consider Problem 2.1.1 in the case where the equality constraint $\mathbf{h}(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^n}$ is active but all inequality constraints are inactive. As explained at the start of Sect. 2.1, this problem corresponds to the case where the cross-section \mathbf{a} and displacement \mathbf{u} of Problem 1.2.2 are $\mathbf{x} = (\mathbf{a}^\top, \mathbf{u}^\top)^\top$, and where the cost function $f_0(\mathbf{u}) = \mathbf{p} \cdot \mathbf{u}$ is written as $f_0(\mathbf{x})$ and the state equation $\mathbf{K}(\mathbf{a})\mathbf{u} - \mathbf{p} = \mathbf{0}_{\mathbb{R}^n}$ is expressed as an equality constraint $\mathbf{h}(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^n}$.

In the case that $n < d$ equality constraints are given, it is considered that n elements become dependent variables. Therefore, putting $\mathbf{u} \in U = \mathbb{R}^n$ and $\xi \in \Xi = \mathbb{R}^{d-n}$ for the remaining elements, we can write $\mathbf{x} = (\xi^\top, \mathbf{u}^\top)^\top \in X = \Xi \times U$. In optimum design problems, ξ is called the design variable. In this section, while being careful that $X = \Xi \times U$, let us consider the following problem. Since f_1, \dots, f_m do not appear in this section, we denote f_0 by f .

Problem 2.6.1 (Optimization Problems with Equality Constraints) Let $X = \mathbb{R}^d$. If $f : X \rightarrow \mathbb{R}$ and $\mathbf{h} = (h_1, \dots, h_n)^\top : X \rightarrow \mathbb{R}^n$ are given with $n < d$, find \mathbf{x} satisfying

$$\min_{\mathbf{x} \in X} \{ f(\mathbf{x}) \mid \mathbf{h}(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^n} \}.$$

□

2.6.1 A Necessary Condition for Local Minimizers

Let us consider a relationship that is established when \mathbf{x} is a local minimizer of Problem 2.6.1. Figure 2.14 illustrates a local minimum when $X = \mathbb{R}^2$ and $n = 1$. Let us consider this figure in order to describe the relationship that is established at local minimums and then consider the general case.

Let \mathbf{x} in Fig. 2.14 designate a local minimum. That is, if we let B_X denote a neighborhood of \mathbf{x} and move \mathbf{x} to $\mathbf{x} + \mathbf{y} \in B_X$ while satisfying the equality constraint $h_1(\mathbf{x} + \mathbf{y}) = 0$, then $f(\mathbf{x}) \leq f(\mathbf{x} + \mathbf{y})$. Here it is assumed that f and h_1 are elements of $C^1(B_X; \mathbb{R})$. Then $\partial_X f = \mathbf{g}$ and $\partial_X h_1$ can be defined and the following relationships hold:

- (1) \mathbf{y} and $\partial_X h_1$ are orthogonal,
- (2) \mathbf{y} and \mathbf{g} are orthogonal.

Relationship (1) expresses that the constraint is satisfied along the direction \mathbf{y} of a point that satisfies the constraint. In fact, we have

$$h_1(\mathbf{x} + \mathbf{y}) = h_1(\mathbf{x}) + \partial_X h_1 \cdot \mathbf{y} + o(\|\mathbf{y}\|_{\mathbb{R}^d}),$$

and if $h_1(\mathbf{x} + \mathbf{y}) = h_1(\mathbf{x})$ at a point $\mathbf{x} + \mathbf{y} \in B_X$ of a neighborhood of \mathbf{x} , then $\partial_X h_1 \cdot \mathbf{y} = 0$. On the other hand, (2) states that the value of the cost function does not change even when it is moved in the direction of the constraint. This relationship is the same as in Eq. (2.5.2), where the variation in the direction of the cost function is limited to \mathbf{y} in Theorem 2.5.2. Let us now generalize these relationships and consider necessary conditions relating local minimizers of optimization problems with equality constraints.

We begin by generalizing relationship (1). Let the admissible set be

$$V = \{ \mathbf{x} \in X \mid \mathbf{h}(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^n} \}. \quad (2.6.1)$$

Figure 2.14 shows a curve of points satisfying $h_1 = 0$, where a neighborhood of $\mathbf{x} \in V$ is denoted by $B_X \subset X$. When $\mathbf{h} \in C^1(B_X; \mathbb{R}^n)$,

$$T_V(\mathbf{x}) = \{ \mathbf{y} \in X \mid \mathbf{h}_{\mathbf{x}^\top}(\mathbf{x}) \mathbf{y} = \mathbf{0}_{\mathbb{R}^n} \} \quad (2.6.2)$$

is called the feasible direction set or the tangent plane at \mathbf{x} . Here, $\mathbf{h}_{\mathbf{x}^\top} = (\partial h_i / \partial x_j) = (\partial_X h_1, \dots, \partial_X h_n)^\top \in \mathbb{R}^{n \times d}$ corresponds to the Jacobi matrix of \mathbf{h} with respect to \mathbf{x} . The rank of this matrix is n (in other words, $\partial_X h_1, \dots, \partial_X h_n$ are all linearly independent) and we remark that the tangent to the curve $h_1 = 0$ at \mathbf{x} is denoted by $T_V(\mathbf{x})$ in Fig. 2.14.

The cases of $n = 1$ and $n = 2$ are easy to imagine when $X = \mathbb{R}^3$. Figure 2.15 shows $T_V(\mathbf{x})$ in these cases. When $n = 1$, the set of points V satisfying $h_1 = 0$

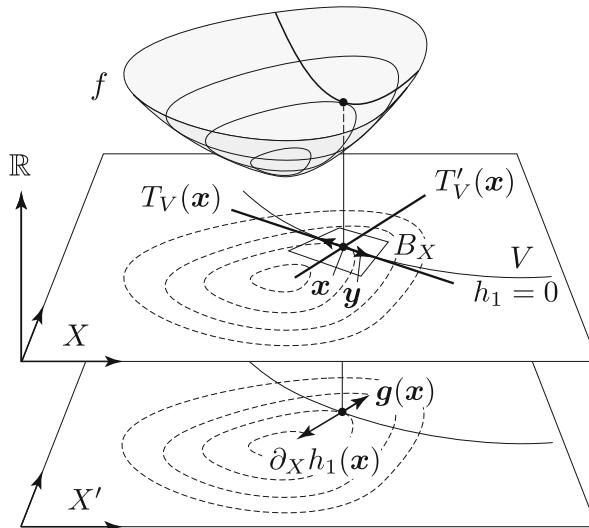


Fig. 2.14 A local minimizer of an optimization problem under an equality constraint ($X = \mathbb{R}^2$, $n = 1$)

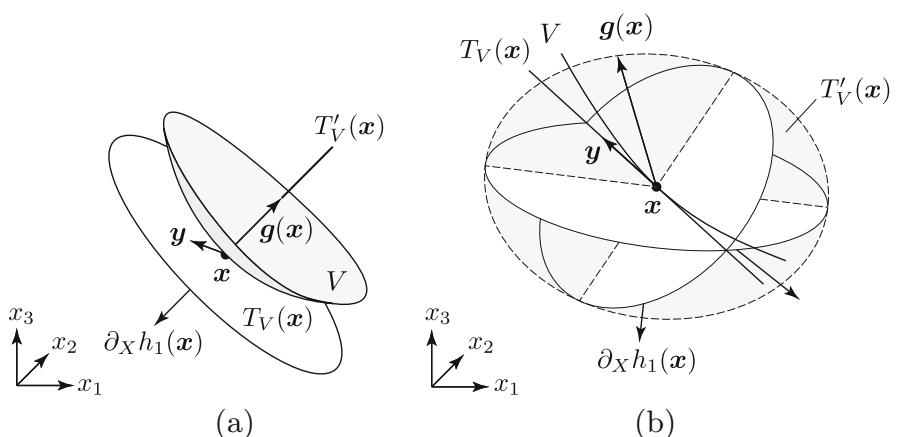


Fig. 2.15 Local minimizers in optimization problems with equality constraints ($X = \mathbb{R}^3$ and $X' = \mathbb{R}^3$ are superimposed). (a) $n = 1$. (b) $n = 2$

forms a surface and $T_V(\mathbf{x})$ is the tangent plane at \mathbf{x} to this surface. When $n = 2$, the set of points V simultaneously satisfying $h_1 = 0$ and $h_2 = 0$ is a curve and $T_V(\mathbf{x})$ is its tangent at \mathbf{x} . When $n = 2$, the fact that the rank of $\mathbf{h}_{\mathbf{x}^\top}(\mathbf{x}) = (\partial_X h_1(\mathbf{x}), \partial_X h_2(\mathbf{x}))^\top$ is $n = 2$ indicates that $\partial_X h_1(\mathbf{x})$ and $\partial_X h_2(\mathbf{x})$ face different directions. Based on Eq. (2.6.2), the definitions of the null space and the image space (also referred to as the kernel space and the range space in Sect. A.3, respectively) allow one to write:

$$T_V(\mathbf{x}) = \text{Ker } \mathbf{h}_{\mathbf{x}^\top}(\mathbf{x}).$$

The generalization of condition (2) is as follows. Condition (2) expresses that \mathbf{g} is orthogonal to all of the vectors contained in $T_V(\mathbf{x})$. We let

$$T'_V(\mathbf{x}) = (T_V(\mathbf{x}))' = \{ \mathbf{z} \in X' \mid \mathbf{z} \cdot \mathbf{y} = 0 \text{ for all } \mathbf{y} \in T_V(\mathbf{x}) \}. \quad (2.6.3)$$

and call $T'_V(\mathbf{x})$ the dual set or the dual plane of $T_V(\mathbf{x})$. Then, if f and \mathbf{h} are first-order differentiable and T'_V can be evaluated, one can obtain the following result.

Theorem 2.6.2 (1st-Order Necessary Conditions for a Local Minimizer) *Let $f \in C^1(X; \mathbb{R})$ and $\mathbf{h} \in C^1(X; \mathbb{R}^n)$ in Problem 2.6.1 and $\partial_X h_1(\mathbf{x}), \dots, \partial_X h_n(\mathbf{x})$ be linearly independent at $\mathbf{x} \in V$. Then if \mathbf{x} is a local minimizer*

$$\mathbf{g}(\mathbf{x}) \cdot \mathbf{y} = 0 \quad (2.6.4)$$

for all $\mathbf{y} \in T_V(\mathbf{x})$. Moreover,

$$\mathbf{g}(\mathbf{x}) \in T'_V(\mathbf{x}). \quad (2.6.5)$$

□

Proof If we let $\mathbf{y} \in T_V(\mathbf{x})$ be arbitrary and suppose that $\mathbf{g}(\mathbf{x}) \cdot \mathbf{y} \neq 0$, then there exists \mathbf{y} such that $\mathbf{g} \cdot \mathbf{y} < 0$. Then a contradiction similar to Theorem 2.5.2 can be obtained and Eq. (2.6.5) is equivalent to Eq. (2.6.4) by the definition of $T'_V(\mathbf{x})$. □

2.6.2 The Lagrange Multiplier Method

Formulated using arbitrary $\mathbf{y} \in T_V(\mathbf{x})$ or $T'_V(\mathbf{x})$, Theorem 2.6.2 expresses a condition that is established when $\mathbf{x} \in V$ is a local minimizer. Nevertheless, even if the theorem's meaning is easy to understand, its evaluation is not necessarily so simple. For this reason, let us consider a method which does not make use of arbitrary $\mathbf{y} \in T_V(\mathbf{x})$ or $T'_V(\mathbf{x})$.

Before making a generalization, let us confirm the fundamental relationship illustrated in Fig. 2.14. Here we assume that both f and h_1 belong to $C^1(B_X; \mathbb{R})$. We also remark that \mathbf{g} is orthogonal to $\mathbf{y} \in T_V(\mathbf{x})$ and that $\mathbf{y} \in T_V(\mathbf{x})$ is also

orthogonal to $\partial_X h_1$. This relationship is equivalent to the fact that \mathbf{g} and $\partial_X h_1$ are oriented in the same direction. This relationship asserts that there exists $\lambda_1 \in \mathbb{R}$ satisfying the following:

$$\mathbf{g} + \lambda_1 \partial_X h_1 = \mathbf{0}_{\mathbb{R}^2}. \quad (2.6.6)$$

In particular, λ_1 satisfying Eq. (2.6.6) cannot exist when \mathbf{g} and $\partial_X h_1$ are non-zero vectors pointing in different directions. The reader is invited to confirm that when two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ are fixed and have different directions, that there does not exist a $\lambda \in \mathbb{R}$ satisfying $\mathbf{a} + \lambda \mathbf{b} = \mathbf{0}_{\mathbb{R}^d}$.

Now let us generalize Eq. (2.6.6). At first, we will generalize the condition:

- $\mathbf{y} \in T_V(\mathbf{x})$ and \mathbf{g} are orthogonal.

This condition can be obtained from the condition that the gradient of cost function f becomes zero when the variable \mathbf{x} moves to the direction where the equality constraints are satisfied. Let V denote the admissible set in Eq. (2.6.1) and assume that the following conditions are satisfied:

Hypothesis 2.6.3 (Implicit Function Theorem Assumptions) Let $d > n$ and $X = \mathbb{R}^d = \Xi \times \mathbb{R}^n$. Also assume that $\mathbf{h} : X \rightarrow \mathbb{R}^n$ satisfies the following conditions in a neighborhood $B_X = B_\Xi \times B_{\mathbb{R}^n}$ of $\mathbf{x} = (\xi_0^\top, \mathbf{u}_0^\top)^\top$:

- (1) $\mathbf{h}(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^n}$ (in other words $\mathbf{x} \in V$),
- (2) $\mathbf{h} \in C^0(B_X; \mathbb{R}^n)$,
- (3) $\mathbf{h}(\xi, \cdot)$ belongs to $C^1(B_{\mathbb{R}^n}; \mathbb{R}^n)$ whenever $\tilde{\mathbf{x}} = (\xi^\top, \mathbf{u}^\top)^\top \in B_\Xi \times B_{\mathbb{R}^n}$,
- (4) the Jacobi matrix $\mathbf{h}_{\mathbf{u}^\top}(\mathbf{x})$ is invertible at \mathbf{x} . □

By the implicit function theorem (Theorem A.4.1), there exists a neighborhood (a convex open set) $U_\Xi \times U_{\mathbb{R}^n} \subset B_\Xi \times B_{\mathbb{R}^n}$ and a continuous function $\mathbf{v} : U_\Xi \rightarrow U_{\mathbb{R}^n}$ (the letter \mathbf{v} is a bold Greek upsilon) and $\mathbf{h}(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^n}$ is equivalent to

$$\mathbf{u} = \mathbf{v}(\xi). \quad (2.6.7)$$

Together with the ξ - \mathbf{u} local coordinate system, Fig. 2.16 also shows a set satisfying the equality constraint when $X = \mathbb{R}^2$ and $n = 1$.

If we set

$$\tilde{\mathbf{x}}(\xi) = (\xi^\top, \mathbf{v}^\top(\xi))^\top,$$

then $\tilde{\mathbf{x}}(\xi) \in V$ and ξ is called the local coordinate of V .

Furthermore, a generalization of

- $\mathbf{y} \in T_V(\mathbf{x})$ and $\partial_X h_1$ are orthogonal.

can be obtained from the conditions that the variable \mathbf{x} satisfies the equality conditions. Let $\tilde{f}(\xi) = f(\tilde{\mathbf{x}}(\xi))$. If $f \in C^1(B_X; \mathbb{R}^n)$, then when $\xi \in B_\Xi$ and

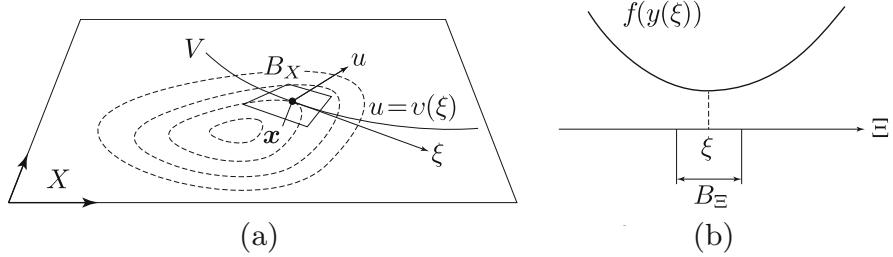


Fig. 2.16 A set satisfying an equality constraint and the $\xi-u$ local coordinate system. (a) $\xi-u$ coordinate system. (b) ξ coordinate system

$\mathbf{x} = \tilde{\mathbf{x}}(\xi) \in B_X$ are local minimizers one has:

$$\partial_{\Xi} \tilde{f}(\xi) = \left(\frac{\partial \tilde{\mathbf{x}}}{\partial \xi^{\top}}(\xi) \right)^{\top} \frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}) = \left(\tilde{\mathbf{x}}_{\xi^{\top}}(\mathbf{x}) \right)^{\top} \mathbf{g}(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^{d-n}}. \quad (2.6.8)$$

The definitions of the null and image spaces allow us to rewrite this relationship as

$$\mathbf{g}(\mathbf{x}) \in T'_V(\mathbf{x}) = \text{Ker} \left(\tilde{\mathbf{x}}_{\xi^{\top}}(\mathbf{x}) \right)^{\top}. \quad (2.6.9)$$

On the other hand, differentiating both sides of $\mathbf{h}(\mathbf{x}) = \mathbf{h}(\tilde{\mathbf{x}}(\xi)) = \mathbf{0}_{\mathbb{R}^n}$ with respect to ξ yields

$$\frac{\partial \mathbf{h}}{\partial \mathbf{x}^{\top}}(\mathbf{x}) \frac{\partial \tilde{\mathbf{x}}}{\partial \xi^{\top}}(\xi) = \mathbf{h}_{\mathbf{x}^{\top}}(\mathbf{x}) \tilde{\mathbf{x}}_{\xi^{\top}}(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^{n \times (d-n)}}. \quad (2.6.10)$$

According to the definitions of the null and image spaces, since the image space of $\tilde{\mathbf{x}}_{\xi^{\top}}(\mathbf{x})$ is the null space of $\mathbf{h}_{\mathbf{x}^{\top}}(\mathbf{x})$, we can write

$$T_V(\mathbf{x}) = \text{Ker} \mathbf{h}_{\mathbf{x}^{\top}}(\mathbf{x}) = \text{Im} \tilde{\mathbf{x}}_{\xi^{\top}}(\mathbf{x}). \quad (2.6.11)$$

These relationships lead us to the following necessary condition for attaining a local minimizer without using $T_V(\mathbf{x})$ or $T'_V(\mathbf{x})$.

Theorem 2.6.4 (1st-Order Necessary Condition for Local Minimizers) Consider Problem 2.6.1 with $f \in C^1(X; \mathbb{R})$ and $\mathbf{h} \in C^1(X; \mathbb{R}^n)$. Let $\partial_X h_1(\mathbf{x}), \dots, \partial_X h_n(\mathbf{x})$ be linearly independent at $\mathbf{x} \in X$. If \mathbf{x} is a local minimizer, then there exists $\lambda \in \mathbb{R}^n$ satisfying

$$\mathbf{g}(\mathbf{x}) + \partial_X \mathbf{h}^{\top}(\mathbf{x}) \lambda = \mathbf{0}_{\mathbb{R}^d}, \quad (2.6.12)$$

$$\mathbf{h}(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^n}. \quad (2.6.13)$$

□

Proof By assumption, Hypothesis 2.6.3 holds at \mathbf{x} . If \mathbf{x} is a local minimizer of Problem 2.6.1, then Eq. (2.6.9) holds. Furthermore, if Eq. (2.6.11) is used, then we have

$$\mathbf{g}(\mathbf{x}) \in T'_V(\mathbf{x}) = (T_V(\mathbf{x}))^\perp = (\text{Im } \tilde{\mathbf{x}}_{\xi^\top}(\mathbf{x}))^\perp = (\text{Ker } \mathbf{h}_{\mathbf{x}^\top}(\mathbf{x}))^\perp$$

and Lemma A.3.1 (relating the orthogonal complement of the null and image spaces) yields

$$\mathbf{g}(\mathbf{x}) \in (\text{Ker } \mathbf{h}_{\mathbf{x}^\top}(\mathbf{x}))^\perp = \text{Im } (\mathbf{h}_{\mathbf{x}^\top}(\mathbf{x}))^\top = \text{Im } (\partial_X h_1, \dots, \partial_X h_n).$$

This relationship is equivalent to Eq. (2.6.12). Moreover, Eq. (2.6.13) holds whenever \mathbf{x} is a local minimizer of Problem 2.6.1. \square

The relation shown at the last part of the proof in Theorem 2.6.4 is a generalization of Eq. (2.6.6). In other words, $\mathbf{g}(\mathbf{x})$ can be given as a linear combination of $\partial_X h_1, \dots, \partial_X h_n$. In Theorem 2.6.4, the vector $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)^\top \in \mathbb{R}^n$ is called a Lagrange multiplier with respect to the equality constraint $\mathbf{h}(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^n}$. Furthermore, Eqs. (2.6.12) and (2.6.13) are called the first-order necessary conditions for the existence of local minimizers under the Lagrange method. The reason for this is because the following relationship holds. The Lagrange function for the optimization problem of Problem 2.6.1 is defined as

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda} \cdot \mathbf{h}(\mathbf{x}). \quad (2.6.14)$$

The derivative of \mathcal{L} with respect to an arbitrary variation $(\mathbf{y}, \hat{\boldsymbol{\lambda}}) \in X \times \mathbb{R}^n$ of $(\mathbf{x}, \boldsymbol{\lambda})$ is

$$\begin{aligned} \mathcal{L}'(\mathbf{x}, \boldsymbol{\lambda}) \left[\mathbf{y}, \hat{\boldsymbol{\lambda}} \right] &= f'(\mathbf{x})[\mathbf{y}] + \boldsymbol{\lambda} \cdot (\partial_{\mathbf{x}^\top} \mathbf{h}(\mathbf{x}) \mathbf{y}) + \hat{\boldsymbol{\lambda}} \cdot \mathbf{h}(\mathbf{x}) \\ &= \mathbf{g}(\mathbf{x}) \cdot \mathbf{y} + (\partial_X \mathbf{h}^\top(\mathbf{x}) \boldsymbol{\lambda}) \cdot \mathbf{y} + \hat{\boldsymbol{\lambda}} \cdot \mathbf{h}(\mathbf{x}). \end{aligned} \quad (2.6.15)$$

Equations (2.6.12) and (2.6.13) of Theorem 2.6.4 are equivalent to the first-order necessary (stationary) condition for the existence of a local minimizer of the Lagrange function, $\mathcal{L}'(\mathbf{x}, \boldsymbol{\lambda}) \left[\mathbf{y}, \hat{\boldsymbol{\lambda}} \right] = 0$ for all $(\mathbf{y}, \hat{\boldsymbol{\lambda}}) \in X \times \mathbb{R}^n$.

We can thus consider using the solution of the following problem as a method for producing candidate solutions to Problem 2.6.1. This method is called the Lagrange multiplier method for optimization problems under an equality constraint.

Problem 2.6.5 (Lagrange Multiplier Method for Equality Constraints) With respect to Problem 2.6.1, let $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ be given by Eq. (2.6.14). Find $(\mathbf{x}, \boldsymbol{\lambda})$ satisfying the stationary condition of $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$:

$$\partial_X \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{g}(\mathbf{x}) + \partial_X \mathbf{h}^\top(\mathbf{x}) \boldsymbol{\lambda} = \mathbf{0}_{\mathbb{R}^d}, \quad (2.6.16)$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}}(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{h}(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^n}. \quad (2.6.17)$$

□

The Lagrange multiplier method will be used in various scenarios going forward. Please note that this method expresses conditions satisfied at local minimizers of Problem 2.6.1, and that it does not directly solve Problem 2.6.1.

Next, let us look at the physical meaning of Lagrange multipliers. Equation (2.6.16) can be written as

$$\lambda_i = -\frac{\left(\mathbf{g}(\mathbf{x}) + \sum_{j \in \{1, \dots, n\}, j \neq i} \lambda_j \partial_X h_j(\mathbf{x})\right) \cdot \mathbf{y}}{\partial_X h_i(\mathbf{x}) \cdot \mathbf{y}}, \quad (2.6.18)$$

where $\mathbf{y} \in X$ is arbitrary. When f and h_1, \dots, h_n are mechanical quantities, λ_i is also a mechanical quantity with units f/h_i . In fact, in Problem 1.1.4, the equality constraint $\mathbf{K}(\mathbf{a})\mathbf{u} = \mathbf{p}$ has the unit of force [N], $f_0 = \mathbf{p} \cdot \mathbf{u}$ has the unit of work [Nm], and \mathbf{v}_0 (introduced as a Lagrange multiplier (adjoint variable) with respect to a state equation) has the unit of displacement [m = Nm/N]. The physical meaning of the Lagrange multiplier method for optimization problems with inequality constraints is also the same. In Problem 1.1.4, f_1 and f_0 had the units of volume [m^3] and work [Nm], respectively. The Lagrange multiplier λ_1 thus has the unit of energy density [$\text{N/m}^2 = \text{Nm/m}^3$].

The cost function in Theorem 2.6.4 was assumed to be first-order differentiable. If it is further assumed to be twice differentiable, then the following results can be obtained. Hereafter, we will write the Hesse matrix with respect to variation of \mathbf{x} of the Lagrange function by $\partial_X \partial_X^\top \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{H}_{\mathcal{L}}(\mathbf{x}, \boldsymbol{\lambda}) \in \mathbb{R}^{d \times d}$.

Theorem 2.6.6 (2nd-Order Necessary Condition) Let $f \in C^2(X; \mathbb{R})$ and $\mathbf{h} \in C^2(X; U)$ in Problem 2.6.1. Also let $\partial_X h_1(\mathbf{x}), \dots, \partial_X h_n(\mathbf{x})$ be linearly independent at $\mathbf{x} \in V$. If \mathbf{x} is a local minimizer

$$\mathbf{y} \cdot (\mathbf{H}_{\mathcal{L}}(\mathbf{x}, \boldsymbol{\lambda}) \mathbf{y}) \geq 0$$

for arbitrary $\mathbf{y} \in T_V(\mathbf{x})$.

□

Proof Calculations similar to Eq. (2.6.8) yield

$$\begin{aligned}\frac{\partial^2 \tilde{f}}{\partial \xi \partial \xi^\top}(\xi) &= \left(\frac{\partial \mathbf{y}}{\partial \xi^\top}(\xi) \right)^\top \frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}^\top}(\mathbf{x}) \frac{\partial \mathbf{y}}{\partial \xi^\top}(\xi) \\ &= \left(\mathbf{y}_{\xi^\top}(\mathbf{x}) \right)^\top \partial_X \partial_X^\top f(\mathbf{x}) \mathbf{y}_{\xi^\top}(\mathbf{x}) \in \mathbb{R}^{(d-n) \times (d-n)}.\end{aligned}$$

If \mathbf{x} is a local minimizer, $\partial^2 \tilde{f} / \partial \xi \partial \xi^\top(\xi)$ is positive definite and Eq. (2.6.10) yields

$$\begin{aligned}\left(\frac{\partial \mathbf{y}}{\partial \xi^\top}(\xi) \right)^\top \frac{\partial h_i}{\partial \mathbf{x} \partial \mathbf{x}^\top}(\mathbf{y}) \frac{\partial \mathbf{y}}{\partial \xi^\top}(\xi) \\ = \left(\mathbf{y}_{\xi^\top}(\mathbf{x}) \right)^\top \partial_X \partial_X^\top h_i(\mathbf{x}) \mathbf{y}_{\xi^\top}(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^{(d-n) \times (d-n)}}.\end{aligned}$$

Since $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda} \cdot \mathbf{h}(\mathbf{x})$, the theorem is established. \square

Based on Theorem 2.6.6, when \mathbf{x} is a local minimizer the Lagrange function $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ can be interpreted as a quadratic approximation of $\tilde{f}(\xi)$ in the tangent plane $T_V(\mathbf{x})$. In fact, if \mathbf{x} is a local minimizer, $\partial \tilde{f} / \partial \xi = \mathbf{0}_{\mathbb{R}^{d-n}}$ and the proof of Theorem 2.6.6 leads to

$$\begin{aligned}\frac{\partial^2 \tilde{f}}{\partial \xi \partial \xi^\top}(\xi) &= \left(\mathbf{y}_{\xi^\top}(\mathbf{x}) \right)^\top \partial_X \partial_X^\top \left(f(\mathbf{x}) + \sum_{i \in \{1, \dots, n\}} \lambda_i h_i \right) \mathbf{y}_{\xi^\top}(\mathbf{x}) \\ &= \left(\mathbf{y}_{\xi^\top}(\mathbf{x}) \right)^\top \partial_X \partial_X^\top \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) \mathbf{y}_{\xi^\top}(\mathbf{x}).\end{aligned}$$

Therefore, if $\boldsymbol{\eta} \in \mathbb{R}^{d-n}$ is arbitrary and we set $\mathbf{y} = \mathbf{y}_{\xi^\top}(\mathbf{x}) \boldsymbol{\eta} \in T_V(\mathbf{x})$ then

$$\tilde{f}(\xi + \boldsymbol{\eta}) = \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) + \mathbf{y}^\top \partial_X \partial_X^\top \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) \mathbf{y} + o\left(\|\mathbf{y}\|_X^2\right).$$

2.6.3 Sufficient Conditions for Local Minimizers

Regarding sufficient conditions for attaining local minimizers, the following results can be obtained.

Theorem 2.6.7 (2nd-Order Sufficient Conditions) *Let $f \in C^2(X; \mathbb{R})$ and $\mathbf{h} \in C^2(X; U)$ in Problem 2.6.5. Also let $\partial_X h_1(\mathbf{x}), \dots, \partial_X h_n(\mathbf{x})$ be linearly*

independent at $\mathbf{x} \in X$. If \mathbf{x} solves Problem 2.6.5, and if there exists a (\mathbf{x}, λ) satisfying

$$\mathbf{y} \cdot (\mathbf{H}_{\mathcal{L}}(\mathbf{x}, \lambda) \mathbf{y}) > 0$$

for arbitrary $\mathbf{y} \in T_V(\mathbf{x})$, then \mathbf{x} is a local minimizer of Problem 2.6.1. \square

Proof Apply the proof of Theorem 2.5.5 to \tilde{f} . \square

2.6.4 An Optimization Problem with an Equality Constraint

Let us consider a spring system and apply the Lagrange multiplier method to solve an optimization problem with an equality constraint.

Exercise 2.6.8 (A Combined Spring Problem) Consider the two-degree-of-freedom spring system shown in Fig. 2.17 and let k_1 and k_2 be positive real constants representing the rigidity of the springs. Also let a be a positive real constant expressing the gap (length) of the spring. Find the displacement $\mathbf{u} = (u_1, u_2)^\top \in \mathbb{R}^2$ at which the potential energy is minimized when the springs are combined. In other words, find \mathbf{u} satisfying

$$\min_{\mathbf{u} \in \mathbb{R}^2} \left\{ f(\mathbf{u}) = \frac{1}{2}k_1u_1^2 + \frac{1}{2}k_2u_2^2 \mid h_1(\mathbf{u}) = a - (u_1 + u_2) = 0 \right\}. \quad \square$$

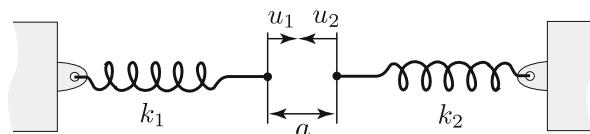
Answer Let us first solve the problem using the substitution method. If we let $u_2 = a - u_1$, then we can write

$$f(\mathbf{u}) = \bar{f}(u_1) = \frac{1}{2}k_1u_1^2 + \frac{1}{2}k_2(a - u_1)^2.$$

Since

$$\frac{d\bar{f}}{du_1}(u_1) = k_1u_1 - k_2(a - u_1) = (k_1 + k_2)u_1 - k_2a = 0,$$

Fig. 2.17 Spring combination problem



the stationary point of \bar{f} becomes

$$u_1 = \frac{k_2}{k_1 + k_2}a, \quad u_2 = a - u_1 = \frac{k_1}{k_1 + k_2}a.$$

Furthermore, since

$$\frac{d^2 \bar{f}}{du_1^2}(u_1) = k_1 + k_2 > 0,$$

$(u_1, u_2)^\top$ is a minimizer.

Next, let us solve the same problem using the Lagrange multiplier method. Let the Lagrange function be

$$\mathcal{L}(\mathbf{u}, \lambda) = \frac{1}{2}k_1u_1^2 + \frac{1}{2}k_2u_2^2 + \lambda(a - u_1 - u_2).$$

The stationary condition for $\mathcal{L}(\mathbf{u}, \lambda)$ becomes

$$\begin{pmatrix} \mathcal{L}_{u_1} \\ \mathcal{L}_{u_2} \\ \mathcal{L}_\lambda \end{pmatrix} = \begin{pmatrix} k_1u_1 - \lambda \\ k_2u_2 - \lambda \\ a - u_1 - u_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

which can be written

$$\begin{pmatrix} k_1 & 0 & -1 \\ 0 & k_2 & -1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ a \end{pmatrix}.$$

The solution of the above is readily obtained:

$$\begin{pmatrix} u_1 \\ u_2 \\ \lambda \end{pmatrix} = \frac{1}{k_1 + k_2} \begin{pmatrix} 1 & -1 & k_2 \\ -1 & 1 & k_1 \\ -k_2 & -k_1 & k_1k_2 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ a \end{pmatrix} = \begin{pmatrix} \frac{k_2}{k_1 + k_2}a \\ \frac{k_1}{k_1 + k_2}a \\ \frac{k_1 + k_2}{k_1k_2}a \end{pmatrix}.$$

We remark that \mathbf{u} agrees with the results from the substitution method and that $\lambda = k_1u_1 = k_2u_2$ carries the meaning of an internal force.

Moreover, the Hesse matrix of the Lagrange function with respect to variation of $\mathbf{u} \in U = \mathbb{R}^2$ is positive definite and independent of \mathbf{u} and λ :

$$\partial_U \partial_U^\top \mathcal{L}(\mathbf{u}, \lambda) = \mathbf{H}_{\mathcal{L}}(\mathbf{u}, \lambda) = \begin{pmatrix} k_1 & 0 \\ 0 & k_2 \end{pmatrix}.$$

By Theorem 2.6.7, \mathbf{u} is a local minimizer. In fact, by Corollary 2.7.10 of Theorem 2.7.9 (shown later), it can be shown that \mathbf{u} is a global minimizer. \square

2.6.5 Direct Differentiation and Adjoint Variable Methods

So far we have investigated necessary and sufficient conditions for the existence of local minimizers with respect to optimization problems under equality constraints (Problem 2.6.1). Let us now replace Problem 2.6.1 with the format of the optimization problem presented in Chap. 1 and look at methods for calculating the derivative of the cost function with respect to variation of the design variable. Chapter 1 referred to these as the direct differentiation method and the adjoint variable method, where only their procedures were considered. Hence we will now define these methods and show the equivalence of the adjoint variable method and the Lagrange multiplier method.

The state determination problem of the optimal design problem presented in Chap. 1 corresponds to an equality constraint. Here, the state determination problem will be defined as follows.

Problem 2.6.9 (Linear System Problem) Let $\Xi = \mathbb{R}^{d-n}$ and $U = \mathbb{R}^n$. Assume that $\mathbf{K} : \Xi \rightarrow \mathbb{R}^{n \times n}$ and $\mathbf{b} : \Xi \rightarrow \mathbb{R}^n$ are given. When $\xi \in \Xi$, find $\mathbf{u} \in U$ satisfying

$$\mathbf{K}(\xi)\mathbf{u} = \mathbf{b}(\xi). \quad (2.6.19)$$

\square

An optimization problem where a state determination problem such as the previous imparts an equality constraint is defined as follows.

Problem 2.6.10 (Optimization Problem with an Equality Constraint) In Problem 2.6.9, let $\mathbf{K} \in C^1(\Xi; \mathbb{R}^{n \times n})$ and $\mathbf{b} \in C^1(\Xi; \mathbb{R}^n)$. When $f \in C^1(\Xi \times U; \mathbb{R})$ is given, find (ξ, \mathbf{u}) which satisfies

$$\min_{(\xi, \mathbf{u}) \in \Xi \times U} \{ f(\xi, \mathbf{u}) \mid \text{Problem 2.6.9} \}. \quad \square$$

Before beginning our explanation, we remark that the derivative of a cost functional (referred to as the cross-sectional derivative in Chap. 1) with respect to variation of a design variable is defined differently than $\mathbf{g}(\mathbf{x})$ from Sect. 2.6. In fact, in Eq. (2.6.8), the derivative of $\tilde{f}(\xi) = f(\xi, \mathbf{v}(\xi))$ with respect to $\xi \in \Xi$ was written $\partial_{\Xi} \tilde{f}(\xi) = (\mathbf{y}_{\xi}^{\top}(\mathbf{x}))^{\top} \mathbf{g}(\mathbf{x})$, where $\mathbf{g}(\mathbf{x})$ was used to refer to $\partial_{\mathbf{x}} f \in \mathbb{R}^d$. On the other hand, $\partial_{\Xi} \tilde{f}_0 \in \mathbb{R}^{d-n}$ was written as \mathbf{g}_0 in Chap. 1. Here, in keeping with the notation in Chap. 2, we will write $\tilde{\mathbf{g}} = \partial_{\Xi} \tilde{f}$.

Let us begin by defining the direct differentiation method. If f , \mathbf{K} and \mathbf{b} are first-order differentiable, then

$$\tilde{\mathbf{g}} = \partial_{\Xi} \tilde{f}(\xi) = \frac{\partial f}{\partial \xi}(\xi, \mathbf{u}) + \left(\frac{\partial \mathbf{u}}{\partial \xi^{\top}}(\xi) \right)^{\top} \frac{\partial f}{\partial \mathbf{u}}(\xi). \quad (2.6.20)$$

On the other hand, the column vector resulting from the partial derivatives of Eq. (2.6.19) with respect to ξ_1, \dots, ξ_{d-n} can be written in a matrix fashion:

$$\frac{\partial \mathbf{K}}{\partial \xi^{\top}} \mathbf{u} + \mathbf{K} \frac{\partial \mathbf{u}}{\partial \xi^{\top}} = \frac{\partial \mathbf{b}}{\partial \xi^{\top}}.$$

In other words we set

$$\frac{\partial \mathbf{K}}{\partial \xi^{\top}} \mathbf{u} = \left(\frac{\partial \mathbf{K}}{\partial \xi_1} \mathbf{u}, \dots, \frac{\partial \mathbf{K}}{\partial \xi_{d-n}} \mathbf{u} \right) \in \mathbb{R}^{n \times (d-n)}. \quad (2.6.21)$$

Therefore

$$\frac{\partial \mathbf{u}}{\partial \xi^{\top}} = \mathbf{K}^{-1} \left(\frac{\partial \mathbf{b}}{\partial \xi^{\top}} - \frac{\partial \mathbf{K}}{\partial \xi^{\top}} \mathbf{u} \right). \quad (2.6.22)$$

The right-hand side of the above equation can be calculated and substituted into Eq. (2.6.20) to obtain $\tilde{\mathbf{g}}$. This method is referred to as the direct differentiation method. Here we view $\partial f / \partial \xi$, $\partial f / \partial \mathbf{u}$, $\partial \mathbf{K} / \partial \xi^{\top}$ and $\partial \mathbf{b} / \partial \xi^{\top}$ as being analytically computable.

In contrast, the adjoint variable method is defined below. First of all, the adjoint problem with respect to f is defined as follows.

Problem 2.6.11 (Adjoint Problem with Respect to f) Let \mathbf{K} and f be given with respect to $\xi \in \Xi$ in Problem 2.6.10. Find $\mathbf{v} \in U$ satisfying

$$\mathbf{K}^{\top} \mathbf{v} = \frac{\partial f}{\partial \mathbf{u}}. \quad (2.6.23)$$

□

The solution \mathbf{v} of Problem 2.6.11 is called an adjoint variable. Combining Eqs. (2.6.22) and (2.6.23) yields

$$\left(\frac{\partial \mathbf{u}}{\partial \xi^{\top}} \right)^{\top} \frac{\partial f}{\partial \mathbf{u}} = \left(\frac{\partial \mathbf{b}}{\partial \xi^{\top}} - \frac{\partial \mathbf{K}}{\partial \xi^{\top}} \mathbf{u} \right)^{\top} \mathbf{K}^{-\top} \mathbf{K}^{\top} \mathbf{v} = \left(\frac{\partial \mathbf{b}}{\partial \xi^{\top}} - \frac{\partial \mathbf{K}}{\partial \xi^{\top}} \mathbf{u} \right)^{\top} \mathbf{v}. \quad (2.6.24)$$

Substituting Eq. (2.6.24) into Eq. (2.6.20) gives

$$\tilde{\mathbf{g}} = \frac{\partial f}{\partial \xi} + \left(\frac{\partial \mathbf{b}}{\partial \xi^\top} - \frac{\partial \mathbf{K}}{\partial \xi^\top} \mathbf{u} \right)^\top \mathbf{v} \in \mathbb{R}^{d-n}. \quad (2.6.25)$$

The method of calculating $\tilde{\mathbf{g}}$ by solving Problem 2.6.11 for \mathbf{v} and using Eq. (2.6.25) is called the adjoint variable method. In this approach, the definition of the Lagrange function is not required. Nevertheless, it can be shown that the Lagrange multiplier method yields the same results as the adjoint variable method.

To see this, let $\mathbf{v} \in \mathbb{R}^n$ be a Lagrange multiplier with respect to an equality constraint (state equation) and define the Lagrange function as

$$\mathcal{L}(\xi, \mathbf{u}, \mathbf{v}) = f(\xi, \mathbf{u}) + \mathbf{v} \cdot (\mathbf{b}(\xi) - \mathbf{K}(\xi) \mathbf{u}).$$

Stationary conditions of $\mathcal{L}(\xi, \mathbf{u}, \mathbf{v})$ with respect to \mathbf{u} and \mathbf{v} are

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{u}}(\xi, \mathbf{u}, \mathbf{v}) &= \frac{\partial f}{\partial \mathbf{u}} - \mathbf{K}^\top \mathbf{v} = \mathbf{0}_{\mathbb{R}^n}, \\ \frac{\partial \mathcal{L}}{\partial \mathbf{v}}(\xi, \mathbf{u}, \mathbf{v}) &= \mathbf{b} - \mathbf{Ku} = \mathbf{0}_{\mathbb{R}^n}. \end{aligned}$$

These agree with Eqs. (2.6.23) and (2.6.19). Moreover, the partial derivative of \mathcal{L} with respect to ξ is consistent with Eq. (2.6.25):

$$\frac{\partial \mathcal{L}}{\partial \xi}(\xi, \mathbf{u}, \mathbf{v}) = \frac{\partial f}{\partial \xi} + \left(\frac{\partial \mathbf{b}}{\partial \xi^\top} - \frac{\partial \mathbf{K}}{\partial \xi^\top} \mathbf{u} \right)^\top \mathbf{v} = \tilde{\mathbf{g}}.$$

From this we see that the Lagrange multiplier method and the adjoint variable method are equivalent. Moreover, the adjoint variable is the same as the Lagrange multiplier.

Moreover, writing the Hessian $\partial_\Xi \partial_\Xi^\top \tilde{f}$ of $\tilde{f}(\xi)$ as $\tilde{\mathbf{H}}$, let us use the Lagrange multiplier method to obtain the $\tilde{\mathbf{H}}$. We define the Lagrange function with respect to $\tilde{f}'(\xi) [\eta_1] = \tilde{\mathbf{g}} \cdot \eta_1$ based on the definition of a Fréchet derivative (Definition 4.5.4) by

$$\mathcal{L}_1(\xi, \mathbf{u}, \mathbf{v}, \mathbf{w}, z)$$

$$= \tilde{\mathbf{g}}(\xi, \mathbf{u}, \mathbf{v}) \cdot \eta_1 + \mathbf{w} \cdot (\mathbf{b}(\xi) - \mathbf{K}(\xi) \mathbf{u}) + z \cdot \left(\frac{\partial f}{\partial \mathbf{u}} - \mathbf{K}^\top \mathbf{v} \right), \quad (2.6.26)$$

where $\mathbf{w} \in U$ and $z \in U$ are the adjoint variables provided for \mathbf{u} and \mathbf{v} in $\tilde{\mathbf{g}}(\xi, \mathbf{u}, \mathbf{v})$. $\eta_1 \in \Xi$ is assumed to be a constant vector in \mathcal{L}_1 . The \mathcal{L}_1 in Eq. (2.6.26) corresponds to \mathcal{L}_{10} in Eq. (1.1.43) with respect to the mean compliance in Chap. 1. When we generalize the argument in Chap. 1, it becomes as follows.

With respect to arbitrary variations $(\eta_2, \hat{u}, \hat{v}, \hat{w}, \hat{z}) \in \Xi \times U^4$ of (ξ, u, v, w, z) , the derivative of \mathcal{L}_1 is written as

$$\begin{aligned} \mathcal{L}'_1(\xi, u, v, w, z) & [\eta_2, \hat{u}, \hat{v}, \hat{w}, \hat{z}] \\ &= \mathcal{L}_{1\xi}(\xi, u, v, w, z) [\eta_2] + \mathcal{L}_{1u}(\xi, u, v, w, z) [\hat{u}] \\ &\quad + \mathcal{L}_{1v}(\xi, u, v, w, z) [\hat{v}] + \mathcal{L}_{1w}(\xi, u, v, w, z) [\hat{w}] \\ &\quad + \mathcal{L}_{1z}(\xi, u, v, w, z) [\hat{z}]. \end{aligned} \quad (2.6.27)$$

The fourth term on the right-hand side of Eq. (2.6.27) vanishes if u is the solution of the state determination problem. If v can be determined as the solution of the adjoint problem, the fifth term of Eq. (2.6.27) also vanishes. Moreover, the second term on the right-hand side of Eq. (2.6.27) is

$$\begin{aligned} & \mathcal{L}_{1u}(\xi, u, v, w, z) [\hat{u}] \\ &= \tilde{g}_{u^\top}(\xi, u, v) [\hat{u}] \cdot \eta_1 - w \cdot (K(\xi) \hat{u}) \\ &= \hat{u} \cdot \left(\tilde{g}_u^\top(\xi, u, v) [\eta_1] - K^\top(\xi) w \right). \end{aligned} \quad (2.6.28)$$

Here, the condition that Eq. (2.6.28) is zero for arbitrary $\hat{u} \in U$ becomes an adjoint problem to determine w . The third term on the right-hand side of Eq. (2.6.27) is

$$\begin{aligned} & \mathcal{L}_{1v}(\xi, u, v, w, z) [\hat{v}] \\ &= \tilde{g}_{v^\top}(\xi, u, v) [\hat{v}] \cdot \eta_1 - z \cdot (K^\top(\xi) \hat{v}) \\ &= \hat{v} \cdot \left(\tilde{g}_v^\top(\xi, u, v) [\eta_1] - K(\xi) z \right). \end{aligned} \quad (2.6.29)$$

Here, the condition that Eq. (2.6.29) is zero for arbitrary $\hat{v} \in U$ corresponds to the adjoint problem for z .

Finally, the first term on the right-hand side of Eq. (2.6.27) becomes

$$\begin{aligned} & \mathcal{L}_{1\xi}(\xi, u, v, w, z) [\eta_2] \\ &= - \left\{ w^\top \left(\frac{\partial K(\xi)}{\partial \xi_1} u \dots \frac{\partial K(\xi)}{\partial \xi_{d-n}} u \right) \right. \\ &\quad \left. + z^\top \left(\frac{\partial K^\top(\xi)}{\partial \xi_1} v \dots \frac{\partial K^\top(\xi)}{\partial \xi_{d-n}} v \right) \right\} \eta_2. \end{aligned}$$

Here, \mathbf{u} , \mathbf{v} , \mathbf{w} ($\boldsymbol{\eta}_1$) and z ($\boldsymbol{\eta}_1$) are assumed to be determined by the conditions above, respectively. If we denote $f(\boldsymbol{\xi}, \mathbf{u})$ here by $\tilde{f}(\boldsymbol{\xi})$, we have the relation:

$$\mathcal{L}_{\mathbf{I}\boldsymbol{\xi}}(\boldsymbol{\xi}, \mathbf{u}, \mathbf{v}, \mathbf{w}, z)[\boldsymbol{\eta}_2] = \tilde{f}''(\boldsymbol{\xi})[\boldsymbol{\eta}_1, \boldsymbol{\eta}_2] = \tilde{\mathbf{g}}_H(\boldsymbol{\xi}, \boldsymbol{\eta}_1) \cdot \boldsymbol{\eta}_2, \quad (2.6.30)$$

where the Hesse gradient $\tilde{\mathbf{g}}_H$ of \tilde{f} is given by

$$\begin{aligned} \tilde{\mathbf{g}}_H(\boldsymbol{\xi}, \boldsymbol{\eta}_1) = & - \left\{ \mathbf{w}^\top(\boldsymbol{\eta}_1) \left(\frac{\partial \mathbf{K}(\boldsymbol{\xi})}{\partial \xi_1} \mathbf{u} \dots \frac{\partial \mathbf{K}(\boldsymbol{\xi})}{\partial \xi_{d-n}} \mathbf{u} \right) \right. \\ & \left. + z^\top(\boldsymbol{\eta}_1) \left(\frac{\partial \mathbf{K}^\top(\boldsymbol{\xi})}{\partial \xi_1} \mathbf{v} \dots \frac{\partial \mathbf{K}^\top(\boldsymbol{\xi})}{\partial \xi_{d-n}} \mathbf{v} \right) \right\}^\top. \end{aligned} \quad (2.6.31)$$

2.6.6 Considerations Relating to the Solution of Optimization Problems with Equality Constraints

Combining the results obtained in this section with a few shown later on allows us to conclude the following about the solution of optimization problems under equality constraints (Problem 2.6.1):

- (1) By Theorem 2.6.4, the solution of the Lagrange multiplier method $(\mathbf{x}, \boldsymbol{\lambda})$ (Problem 2.6.5) satisfies the necessary conditions for a local minimizer. Such \mathbf{x} 's are candidates for local minimizers.
- (2) When $(\mathbf{x}, \boldsymbol{\lambda})$ is a solution of the Lagrange multiplier method (Problem 2.6.5) whose Hesse matrix $\partial_{\mathbf{x}} \partial_{\mathbf{x}}^\top \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{H}_{\mathcal{L}}(\mathbf{x}, \boldsymbol{\lambda})$ (with respect to variation of \mathbf{x} of the Lagrange function) satisfies $\mathbf{y} \cdot (\mathbf{H}_{\mathcal{L}}(\mathbf{x}, \boldsymbol{\lambda}) \mathbf{y}) > 0$ for arbitrary variations $\mathbf{y} \in T_V(\mathbf{x})$ satisfying the equality constraints, then \mathbf{x} yields a local minimum by Theorem 2.6.7.
- (3) Based on Corollary 2.7.10 of Theorem 2.7.9 shown later, when a convex optimization problem is subject to an equality constraint (Problem 2.6.1), the solution \mathbf{x} of the Lagrange multiplier method is the minimizer.
- (4) Based on Corollary 2.7.3 of Theorem 2.7.2 (shown later), even when non-convex optimization problems include equality constraints (Problem 2.6.1), if \tilde{f} is convex then the stationary point of \tilde{f} (that is, the \mathbf{x} for which $\tilde{\mathbf{g}} = \mathbf{0}_{\mathbb{R}^{d-n}}$) can be shown to yield the minimum.

2.7 Optimization Problems Under Inequality Constraints

Let us now change the constraint condition from an equality to an inequality. We will only consider the case when the inequalities are assumed to be such as those presented in Problem 2.1.2 and Problem 2.1.3. Figure 2.18 shows a local minimizer

Fig. 2.18 A local minimizer of an optimization problem under an inequality constraint when $X = \mathbb{R}^2$ and $m = 1$

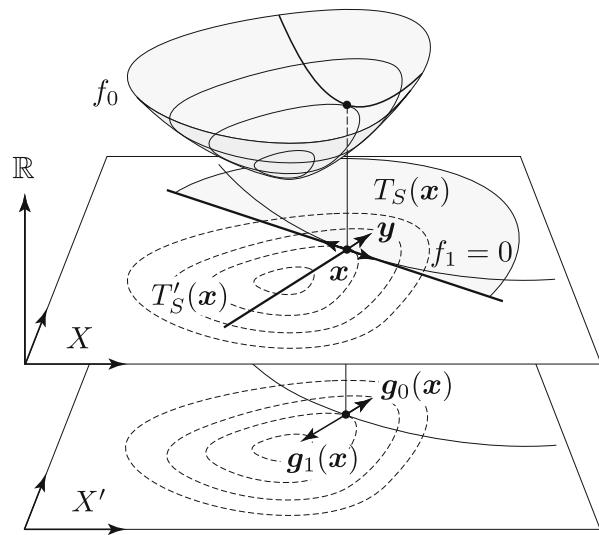
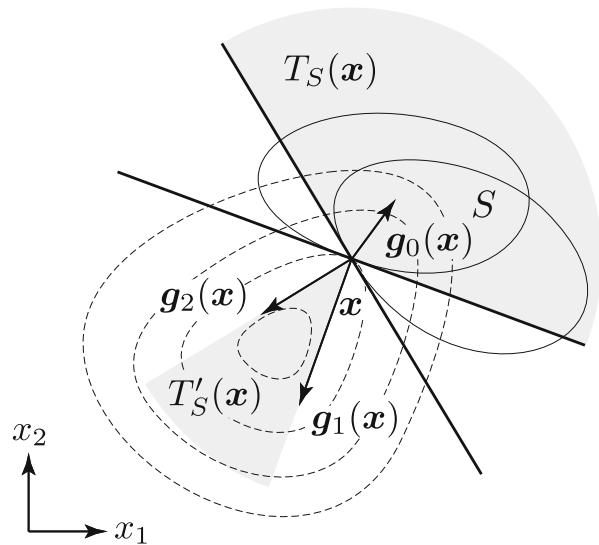


Fig. 2.19 A local minimizer of an optimization problem under an inequality constraint when $X = \mathbb{R}^2$ and $m = 2$



in the case $X = \mathbb{R}^2$ and $m = 1$. When $X = \mathbb{R}^2$ and $m = 2$, the situation is as shown in Fig. 2.19. Using these diagrams, let us first take a look at conditions established at local minimizers and then treat the general case.

2.7.1 Necessary Conditions at Local Minimizers

We will begin with the case illustrated in Fig. 2.18, where one of the inequality constraints is active at the local minimizer \mathbf{x} . In this case, the following can be said:

- (1) The directions \mathbf{y} in which variations of \mathbf{x} are permitted satisfy $\mathbf{g}_1 \cdot \mathbf{y} \leq 0$. These directions are shown in the semicircle region of the figure, where they are denoted by $T_S(\mathbf{x})$.
- (2) If \mathbf{x} is a local minimizer, then f_0 cannot fluctuate and should increase with respect to variations in all directions $\mathbf{y} \in T_S(\mathbf{x})$. This relationship indicates that $\mathbf{g}_0 \cdot \mathbf{y} \geq 0$ for all $\mathbf{y} \in T_S(\mathbf{x})$. Such directions \mathbf{z} satisfying $\mathbf{z} \cdot \mathbf{y} \leq 0$ for all $\mathbf{y} \in T_S(\mathbf{x})$ are shown in the region $T'_S(\mathbf{x})$ of the diagram.
- (3) If \mathbf{x} is a local minimizer, the fact that $\mathbf{g}_0 \cdot \mathbf{y} \geq 0$ holds for all $\mathbf{y} \in T_S(\mathbf{x})$ is equivalent to $-\mathbf{g}_0$ being included in $T'_S(\mathbf{x})$.

When two inequality constraints are active, the situation becomes as is shown in Fig. 2.19. The set of directions $T_S(\mathbf{x})$ in which variations from \mathbf{x} are permissible is wedge-shaped because there are two inequality constraints which must be satisfied simultaneously. In response to this, $T'_S(\mathbf{x})$ is broader than when there is just one inequality constraint and its region becomes wedge-shaped. When \mathbf{x} is a local minimizer, then $\mathbf{g}_0 \cdot \mathbf{y} \geq 0$ holds for all $\mathbf{y} \in T_S(\mathbf{x})$, as does the fact that $-\mathbf{g}_0$ is included in $T'_S(\mathbf{x})$. Figure 2.20 shows the state of a local minimizer when $X = \mathbb{R}^3$.

We now turn to generalizing the above observations. In Figs. 2.18, 2.19, 2.20, $T_S(\mathbf{x})$ and $T'_S(\mathbf{x})$ were defined using \mathbf{g}_1 and \mathbf{g}_2 . Here we define $C_S(\mathbf{x})$ to be the set of admissible directions including $T_S(\mathbf{x})$, and conduct a similar discussion centering on $C_S(\mathbf{x})$ and $C'_S(\mathbf{x})$. Moreover, the relationships between $T_S(\mathbf{x})$ and $C_S(\mathbf{x})$ can

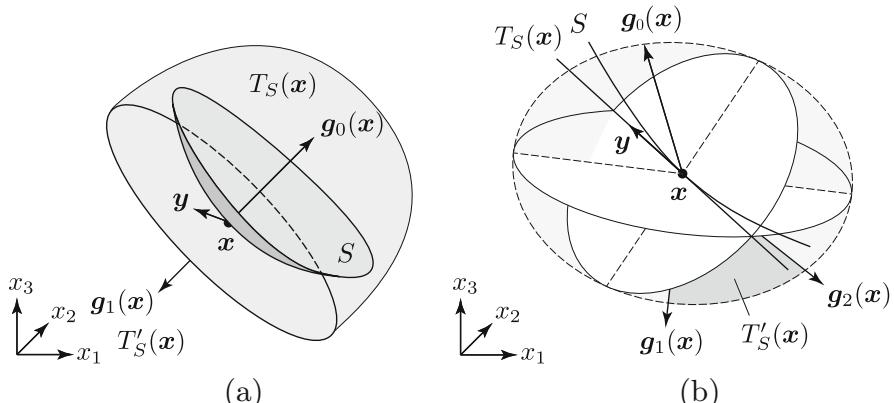


Fig. 2.20 A local minimizer in an optimization problem with an inequality constraint ($X = \mathbb{R}^3$ and $X' = \mathbb{R}^3$ are shown superimposed.) (a) $m = 1$. (b) $m = 2$

be viewed equivalently when the conditions of Proposition 2.7.4 (presented later) are satisfied.

The set S of admissible design variables satisfying the inequality constraints are defined as in Eq. (2.1.1). Given $\mathbf{x} \in S$, the active constraints are indicated by the following set:

$$I_A(\mathbf{x}) = \{i \in \{1, \dots, m\} \mid f_i(\mathbf{x}) = 0\} = \{i_1, \dots, i_{|I_A(\mathbf{x})|}\}. \quad (2.7.1)$$

Considering the set of sequences $\{\mathbf{y}_k\}_{k \in \mathbb{N}} \in S$ converging to $\mathbf{x} \in S$ and having a direction \mathbf{y} , the set

$$C_S(\mathbf{x}) = \left\{ \mathbf{y} \in X \mid \frac{\mathbf{y}}{\|\mathbf{y}\|} = \lim_{k \rightarrow \infty} \frac{\mathbf{y}_k - \mathbf{x}}{\|\mathbf{y}_k - \mathbf{x}\|} \text{ for } \mathbf{y} \neq \mathbf{0}_X \right\}$$

is called the feasible direction set or the tangential cone of S . The dual cone of $C_S(\mathbf{x})$ is the set:

$$C'_S(\mathbf{x}) = \{z \in X' \mid z \cdot \mathbf{y} \leq 0 \text{ for all } \mathbf{y} \in C_S(\mathbf{x})\}.$$

Our next result follows easily from the above considerations when C'_S can be evaluated and f_0 is first-order differentiable.

Theorem 2.7.1 (1st-Order Necessary Condition for Local Minimizers) *Let $f_0 \in C^1(X; \mathbb{R})$ in Problem 2.1.2. If \mathbf{x} is a local minimizer, then for arbitrary $\mathbf{y} \in C_S(\mathbf{x})$,*

$$\mathbf{g}_0(\mathbf{x}) \cdot \mathbf{y} \geq 0. \quad (2.7.2)$$

Moreover,

$$-\mathbf{g}_0(\mathbf{x}) \in C'_S(\mathbf{x}). \quad (2.7.3)$$

□

Proof If we suppose $\mathbf{g}_0(\mathbf{x}) \cdot \mathbf{y} \neq 0$ for all $\mathbf{y} \in C_S(\mathbf{x})$, then there exists \mathbf{y} such that $\mathbf{g}_0(\mathbf{x}) \cdot \mathbf{y} < 0$. The same contradiction as was obtained in the proof of Theorem 2.5.2 can then be obtained. Moreover, Eq. (2.7.3) is equivalent to Eq. (2.7.2). □

2.7.2 Necessary and Sufficient Conditions for Global Minimizers

If Problem 2.1.2 is a convex optimization problem and C'_S can be evaluated, then the following necessary and sufficient condition satisfied by global minimizers can be obtained.

Theorem 2.7.2 (1st-Order Necessary and Sufficient Condition) *In Problem 2.1.2, let f_0 be an element of $C^1(X; \mathbb{R})$, f_1, \dots, f_m be elements of $C^0(X; \mathbb{R})$, and f_0, \dots, f_m be convex functions. Also let S be given by Eq. (2.1.1). Then the following condition is both necessary and sufficient for $\mathbf{x} \in S$ to be a global minimizer:*

$$-\mathbf{g}_0(\mathbf{x}) \in C'_S(\mathbf{x}). \quad \square$$

Proof Necessity follows directly from Theorem 2.7.1 and so we only show the sufficient condition. Let each member of the sequence $\{\beta_k\}_{k \in \mathbb{N}}$ satisfy $\beta_k \in (0, 1)$ and $\beta_k \rightarrow 0$. Given $\mathbf{y} \in S$, construct $\{\mathbf{y}_k\}_{k \in \mathbb{N}}$ such that $\mathbf{y}_k = (1 - \beta_k)\mathbf{x} + \beta_k\mathbf{y}$. Since S is convex, $\{\mathbf{y}_k\}_{k \in \mathbb{N}} \subseteq S$. By the definition of $C_S(\mathbf{x})$, it follows that $\mathbf{y} - \mathbf{x} \in C_S(\mathbf{x})$. Therefore, by the definition of $C'_S(\mathbf{x})$, and since $-\mathbf{g}_0(\mathbf{x}) \in C'_S(\mathbf{x})$,

$$-\mathbf{g}_0(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}) \leq 0.$$

Since f_0 is a convex function, Theorem 2.4.4 implies that

$$\mathbf{g}_0(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}) \leq f_0(\mathbf{y}) - f_0(\mathbf{x}).$$

Hence, $f_0(\mathbf{x}) \leq f_0(\mathbf{y})$. \square

If the global minimizer occurs at an interior point of S , Theorem 2.7.2 is as follows (this result is equivalent to Theorem 2.5.6).

Corollary 2.7.3 (1st-Order Necessary and Sufficient Condition) *In Problem 2.1.2, let f_0 be $C^1(X; \mathbb{R})$ and f_1, \dots, f_m be convex functions belonging to $C^0(X; \mathbb{R})$. Let S be given by Eq. (2.1.1). Then the following is both necessary and sufficient for an internal point \mathbf{x} of S to yield the global minimum in Problem 2.1.3:*

$$\mathbf{g}_0(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^d}. \quad \square$$

Proof If \mathbf{x} is an internal point of S , $C_S(\mathbf{x}) = X$. Hence, we obtain $C'_S(\mathbf{x}) = \{\mathbf{0}_{\mathbb{R}^d}\}$. Theorem 2.7.2 then implies $\mathbf{g}_0(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^d}$. \square

2.7.3 KKT Conditions

Conditions governing all \mathbf{y} in $C_S(\mathbf{x})$ or C'_S were included in the necessary and sufficient conditions for the existence of local and global minimizers of Problem 2.1.2. However, checking such conditions is not necessarily easy. Therefore, as we did in the case of optimization problems with equality constraints, we will consider expressions using Lagrange functions here as well.

Let us first consider the situation shown in Fig. 2.18, where one inequality constraint is active. Since the inequality constraint condition is active, even if there is an equality constraint, \mathbf{x} will yield a local minimum. Then Eq. (2.6.6) (which was used with an equality constraint) can be rewritten as

$$\mathbf{g}_0 + \lambda_1 \mathbf{g}_1 = \mathbf{0}_{\mathbb{R}^2}. \quad (2.7.4)$$

However, the range of admissible variations is enlarged in the presence of inequality constraints. Let us consider this in detail. Taking the inner product of Eq. (2.7.4) with an arbitrary $\mathbf{y} \in \mathbb{R}^2$ yields

$$\mathbf{g}_0 \cdot \mathbf{y} + \lambda_1 \mathbf{g}_1 \cdot \mathbf{y} = 0. \quad (2.7.5)$$

If \mathbf{y} is a direction in which the inequality constraint is satisfied, then $\mathbf{g}_1 \cdot \mathbf{y} \leq 0$. Moreover, if \mathbf{x} is a local minimizer, then the cost function remains constant or increases for such a \mathbf{y} and we obtain $\mathbf{g}_0 \cdot \mathbf{y} \geq 0$. In order to simultaneously satisfy these two conditions, we require:

$$\lambda_1 \geq 0. \quad (2.7.6)$$

Moreover, the original inequality constraint is satisfied at the local minimizer

$$f_1 \leq 0. \quad (2.7.7)$$

Furthermore, when the inequality constraint is inactive ($f_1(\mathbf{x}) < 0$), since this situation is the same as when there are no inequality constraints, $\lambda_1 = 0$. On the other hand, when the inequality constraint is active ($f_1(\mathbf{x}) = 0$), Eq. (2.7.6) is established. These relationships are satisfied if

$$\lambda_1 f_1 = 0. \quad (2.7.8)$$

Equations (2.7.4), (2.7.6), (2.7.8) and (2.7.7) are the conditions established at local minimizers when there is one active inequality constraint. These conditions correspond to the KKT conditions with $m = 1$ (described later).

Next we consider the case when two inequality constraints are active at a local minimizer, such as is shown in Fig. 2.19. In this case as well, imposing equality constraints is equivalent to the existence of certain $\lambda_1, \lambda_2 \in \mathbb{R}$ satisfying

$$\mathbf{g}_0 + \lambda_1 \mathbf{g}_1 + \lambda_2 \mathbf{g}_2 = \mathbf{0}_{\mathbb{R}^2}. \quad (2.7.9)$$

Let us rewrite Eq. (2.7.9) as

$$-\mathbf{g}_0 = \lambda_1 \mathbf{g}_1 + \lambda_2 \mathbf{g}_2. \quad (2.7.10)$$

If we fix \mathbf{g}_1 and \mathbf{g}_2 and take

$$\lambda_1 \geq 0, \quad \lambda_2 \geq 0, \quad (2.7.11)$$

then the vector on the right-hand side of Eq. (2.7.10) expresses the region drawn as $T'_S(\mathbf{x})$ in Fig. 2.19 (the definition of $T'_S(\mathbf{x})$ is given later as Eq. (2.7.15)). Equation (2.7.10) is therefore a condition for which $C'_S(\mathbf{x})$ of Theorem 2.7.1 is rewritten as $T'_S(\mathbf{x})$. Moreover, at local minimizers the original inequality constraints are satisfied:

$$f_1 \leq 0, \quad f_2 \leq 0. \quad (2.7.12)$$

For the reasons explained above, the following equations hold at local minimizers:

$$\lambda_1 f_1 = 0, \quad \lambda_2 f_2 = 0. \quad (2.7.13)$$

Therefore, Eqs. (2.7.9), (2.7.11), (2.7.13) and (2.7.12) are the conditions holding at local minimizers when two inequality constraints are active. These are the KKT conditions with $m = 2$ (described later).

In order to generalize the above results we now state a few required definitions and assumptions. Let a neighborhood of $\mathbf{x} \in S$ be denoted by $B_X \subset X$. Given $i \in I_A(\mathbf{x})$ and $f_i \in C^1(B_X; \mathbb{R})$, let $\mathbf{g}_i(\mathbf{y})$ be linearly independent with respect to $\mathbf{y} \in B_X$. Then the linearized feasible direction set at \mathbf{x} is the set

$$T_S(\mathbf{x}) = \{ \mathbf{y} \in X \mid \mathbf{g}_i(\mathbf{x}) \cdot \mathbf{y} \leq 0 \text{ for all } i \in I_A(\mathbf{x}) \}.$$

Corresponding to the null space in the optimization problem with an equality constraint, $T_S(\mathbf{x})$ will be written as

$$T_S(\mathbf{x}) = \text{Kco}(\mathbf{g}_{i_1}(\mathbf{x}), \dots, \mathbf{g}_{i_k}(\mathbf{x}))^\top. \quad (2.7.14)$$

Moreover,

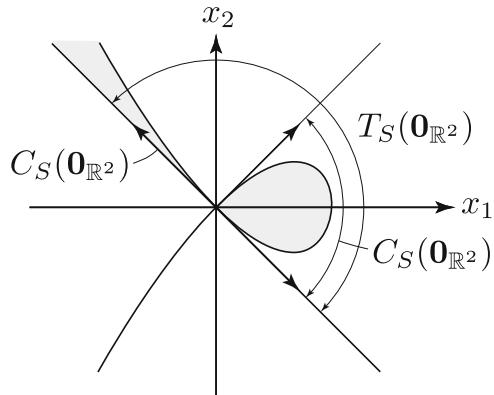
$$T'_S(\mathbf{x}) = \{ \mathbf{z} \in X' \mid \mathbf{z} \cdot \mathbf{y} \leq 0 \text{ for all } \mathbf{y} \in T_S(\mathbf{x}) \} \quad (2.7.15)$$

is called the dual cone of $T_S(\mathbf{x})$.

Let us now take a look at the difference between $T_S(\mathbf{x})$ and $C_S(\mathbf{x})$. We note that $T_S(\mathbf{x})$ is a closed convex polyhedral cone, but that $C_S(\mathbf{x})$ need not share this property [50]. For example, when

$$S = \left\{ \mathbf{y} \in \mathbb{R}^2 \mid f_1 = -y_1^2 + y_1^3 + y_2^2 \leq 0, \quad f_2 = -y_1 - y_2 \leq 0 \right\}, \quad (2.7.16)$$

Fig. 2.21 An example when C_S is not a closed convex polyhedral cone: $C_S(\mathbf{0}_{\mathbb{R}^2})$



we obtain

$$\begin{aligned} C_S(\mathbf{0}_{\mathbb{R}^2}) &= \left\{ \mathbf{y} \in \mathbb{R}^2 \mid y_1 + y_2 \geq 0, y_1 - y_2 \geq 0 \right\} \\ &\cup \left\{ \mathbf{y} \in \mathbb{R}^2 \mid y_1 + y_2 = 0 \right\}, \\ C'_S(\mathbf{0}_{\mathbb{R}^2}) &= \left\{ \alpha (-1, -1)^\top \in \mathbb{R}^2 \mid \alpha \geq 0 \right\}. \end{aligned}$$

Figure 2.21 shows $C_S(\mathbf{0}_{\mathbb{R}^2})$ which is clearly not a closed convex polyhedral cone. In general, given $\mathbf{x} \in S$ we have

$$C_S(\mathbf{x}) \subseteq T_S(\mathbf{x}).$$

For example, when S is given by Eq. (2.7.16), we obtain

$$\begin{aligned} T_S(\mathbf{0}_{\mathbb{R}^2}) &= \left\{ \mathbf{y} \in \mathbb{R}^2 \mid y_1 + y_2 \geq 0 \right\}, \\ T'_S(\mathbf{0}_{\mathbb{R}^2}) &= \left\{ \alpha (-1, -1)^\top \in \mathbb{R}^2 \mid \alpha \geq 0 \right\}. \end{aligned}$$

Sufficient conditions for establishing the equality $T_S(\mathbf{x}) = C_S(\mathbf{x})$ are called first-order constraint qualifications. Cottle's constraint qualification, shown next, is one such condition [51].

Proposition 2.7.4 (Cottle's Constraint Qualification) *Let S be given by Eq. (2.1.1) in Problem 2.1.2. Also, let $\mathbf{x} \in S$ and $I_A(\mathbf{x})$ be given by Eq. (2.7.1). When $\mathbf{g}_i(\mathbf{x})$ is linear for all $i \in I_A(\mathbf{x})$, if there exists $\mathbf{y} \in X$ such that $\mathbf{g}_i(\mathbf{x}) \cdot \mathbf{y} \leq 0$, one has $T_S(\mathbf{x}) = C_S(\mathbf{x})$. In the case that some $\mathbf{g}_i(\mathbf{x})$ is nonlinear, if there exists $\mathbf{y} \in X$ such that $\mathbf{g}_i(\mathbf{x}) \cdot \mathbf{y} < 0$, one has $T_S(\mathbf{x}) = C_S(\mathbf{x})$. \square*

If linear constraint qualifications such as these are used, the conditions holding at local minimizers of Problem 2.1.2 can be expressed in the following way.

Theorem 2.7.5 (KKT Conditions) *In Problem 2.1.2, let f_0, \dots, f_m be elements of $C^1(X; \mathbb{R})$. Given $\mathbf{x} \in S$, let the linear constraint qualification be satisfied. If \mathbf{x} is a local minimizer, then there exists $(\lambda_1, \dots, \lambda_m)^\top \in \mathbb{R}^m$ satisfying*

$$\mathbf{g}_0(\mathbf{x}) + \sum_{i \in \{1, \dots, m\}} \lambda_i \mathbf{g}_i(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^d}, \quad (2.7.17)$$

$$f_i(\mathbf{x}) \leq 0 \quad \text{for } i \in \{1, \dots, m\}, \quad (2.7.18)$$

$$\lambda_i f_i(\mathbf{x}) = 0 \quad \text{for } i \in \{1, \dots, m\}, \quad (2.7.19)$$

$$\lambda_i \geq 0 \quad \text{for } i \in \{1, \dots, m\}. \quad (2.7.20)$$

□

Proof Given arbitrary $\mathbf{t} = (t_1, \dots, t_m)^\top \in \mathbb{R}^m$, the inequality constraint of Problem 2.1.2 can be written as

$$h_i(\mathbf{x}, t_i) = f_i(\mathbf{x}) + t_i^2 = 0 \quad \text{for } i \in \{1, \dots, m\}. \quad (2.7.21)$$

Include \mathbf{t} in the design variables and let the Lagrange function for Problem 2.1.2 be given by

$$\mathcal{L}(\mathbf{x}, \mathbf{t}, \boldsymbol{\lambda}) = f_0(\mathbf{x}) + \sum_{i \in \{1, \dots, m\}} \lambda_i h_i(\mathbf{x}, t_i). \quad (2.7.22)$$

When (\mathbf{x}, \mathbf{t}) is a local minimizer of f_0 which satisfies the equality constraint Eq. (2.7.21), then by Theorem 2.6.4 the following hold:

$$\mathcal{L}_x(\mathbf{x}, \mathbf{t}, \boldsymbol{\lambda}) = \mathbf{g}_0(\mathbf{x}) + \sum_{i \in \{1, \dots, m\}} \lambda_i \mathbf{g}_i(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^d}, \quad (2.7.23)$$

$$\mathcal{L}_{t_i}(\mathbf{x}, \mathbf{t}, \boldsymbol{\lambda}) = 2\lambda_i t_i = 0 \quad \text{for } i \in \{1, \dots, m\}, \quad (2.7.24)$$

$$\mathcal{L}_{\lambda_i}(\mathbf{x}, \mathbf{t}, \boldsymbol{\lambda}) = f_i(\mathbf{x}) + t_i^2 = 0 \quad \text{for } i \in \{1, \dots, m\}. \quad (2.7.25)$$

Equations (2.7.23) and (2.7.25) are equivalent to Eqs. (2.7.17) and (2.7.18), respectively. Moreover, Eq. (2.7.19) can be obtained by multiplying both sides of Eq. (2.7.24) by t_i and using Eq. (2.7.21).

The fact that Eq. (2.7.20) holds can be confirmed as follows. If \mathbf{x} is a local minimizer of Problem 2.1.2 and the linear constraint qualification is satisfied, then by Theorem 2.7.1, Eq. (2.7.14) and Farkas's lemma (Lemma A.3.2)

$$\begin{aligned} -\mathbf{g}_0(\mathbf{x}) &\in C'_S(\mathbf{x}) = T'_S(\mathbf{x}) = (T_S(\mathbf{x}))' = \left(\text{Kco}(\mathbf{g}_{i_1}(\mathbf{x}), \dots, \mathbf{g}_{i_k}(\mathbf{x}))^\top \right)' \\ &= \text{Ico}(\mathbf{g}_{i_1}, \dots, \mathbf{g}_{i_k}), \end{aligned}$$

where $i_1, \dots, i_k \in I_A(\mathbf{x})$. Here we have written

$$\text{Ico}(\mathbf{g}_{i_1}, \dots, \mathbf{g}_{i_k}) = \{(\mathbf{g}_{i_1}, \dots, \mathbf{g}_{i_k}) \lambda \in X' \mid \lambda \geq \mathbf{0}_{\mathbb{R}^k}\},$$

where $\lambda = (\lambda_{i_1}, \dots, \lambda_{i_k})^\top$. Also $\lambda_i = 0$ when $i \notin I_A(\mathbf{x})$. This relationship shows that Eq. (2.7.20) holds. \square

Equations (2.7.17) to (2.7.20) are called the Karush–Kuhn–Tucker (KKT) conditions. Equation (2.7.18) states that \mathbf{x} satisfies the inequality constraints. Equation (2.7.19) is called a complementarity condition and has the effect of removing \mathbf{g}_i from Eq. (2.7.17) by setting $\lambda_i = 0$ with respect to an inactive constraint $f_i(\mathbf{x}) < 0$. Finally, as described in the considerations used in Fig. 2.19, the conditions which combine Eqs. (2.7.17) and (2.7.20) are those establishing when $-\mathbf{g}_0$ is contained in $T'_S(\mathbf{x})$, and are the conditions that allow one to rewrite $C'_S(\mathbf{x})$ (Theorem 2.7.1) as $T'_S(\mathbf{x})$. The variable t is called a slack variable.

Let us also define the Lagrange function approach for optimization problems under inequality constraints. Its relationship with the duality theorem (shown later) will be considered, and we set

$$\mathcal{L}(\mathbf{x}, \lambda) = \begin{cases} f_0(\mathbf{x}) + \sum_{i \in \{1, \dots, m\}} \lambda_i f_i(\mathbf{x}) & (\lambda \geq \mathbf{0}_{\mathbb{R}^m}), \\ -\infty & (\lambda \not\geq \mathbf{0}_{\mathbb{R}^m}). \end{cases} \quad (2.7.26)$$

Here, $\lambda = (\lambda_1, \dots, \lambda_m)^\top \in \mathbb{R}^m$ is a Lagrange multiplier and we remark that Eqs. (2.7.17) and (2.7.18) can be rewritten in terms of $\mathcal{L}(\mathbf{x}, \lambda)$. The method of using solutions of the next problem as candidates for solutions to Problem 2.1.2 is called the Lagrange multiplier method for an optimization problem with an inequality constraint.

Problem 2.7.6 (Lagrange Multiplier Method for Inequality Constraints) Let $\mathcal{L}(\mathbf{x}, \lambda)$ be given by Eq. (2.7.26) in Problem 2.1.2. Find (\mathbf{x}, λ) which satisfies the KKT conditions

$$\mathcal{L}_x(\mathbf{x}, \lambda) = \mathbf{g}_0(\mathbf{x}) + \sum_{i \in \{1, \dots, m\}} \lambda_i \mathbf{g}_i(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^d}, \quad (2.7.27)$$

$$\mathcal{L}_{\lambda_i}(\mathbf{x}, \lambda) = f_i(\mathbf{x}) \leq 0 \quad \text{for } i \in \{1, \dots, m\}, \quad (2.7.28)$$

$$\lambda_i f_i(\mathbf{x}) = 0 \quad \text{for } i \in \{1, \dots, m\}, \quad (2.7.29)$$

$$\lambda_i \geq 0 \quad \text{for } i \in \{1, \dots, m\}. \quad (2.7.30)$$

\square

We can obtain the following necessary condition holding at local minimizers of Problem 2.1.2. With respect to $(\mathbf{x}, \boldsymbol{\lambda}) \in X \times \mathbb{R}^m$ satisfying the KKT conditions, in addition to $I_A(\mathbf{x})$ and S of Eqs. (2.7.1) and (2.1.1) respectively, we define

$$\bar{I}_A(\boldsymbol{\lambda}) = \{i \in \{1, \dots, m\} \mid \lambda_i > 0\}, \quad (2.7.31)$$

$$\bar{S} = S \cap \{\mathbf{x} \in X \mid f_i(\mathbf{x}) = 0, i \in \bar{I}_A(\boldsymbol{\lambda})\} \quad (2.7.32)$$

and $T_{\bar{S}}(\mathbf{x})$ as the feasible direction set of \bar{S} at \mathbf{x} . Moreover, we write the Hesse matrix with respect to \mathbf{x} of the Lagrange function $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ by $\mathbf{H}_{\mathcal{L}}(\mathbf{x}, \boldsymbol{\lambda}) = \partial_{\mathbf{x}} \partial_{\mathbf{x}}^{\top} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$.

Theorem 2.7.7 (2nd-Order Necessary Condition for Local Minimizers) *Let f_0, \dots, f_m be elements of $C^2(X; \mathbb{R})$ in Problem 2.1.2. Given $\mathbf{x} \in X$, assume that \mathbf{g}_i is linearly independent with respect to $i \in I_A(\mathbf{x})$ and satisfies the linear constraint qualification. In this case, if \mathbf{x} is a local minimizer of Problem 2.1.2, then the following holds for an arbitrary tangential vector $\mathbf{y} \in T_{\bar{S}}(\mathbf{x})$:*

$$\mathbf{y} \cdot (\mathbf{H}_{\mathcal{L}}(\mathbf{x}, \boldsymbol{\lambda}) \mathbf{y}) \geq 0. \quad \square$$

Proof Let $\{\mathbf{y}_k\}_{k \in \mathbb{N}} \in \bar{S}$ satisfy $\lambda_i f_i(\mathbf{y}_k) = 0$ for each $i \in \{1, \dots, m\}$ and $t / \|t\|_{\mathbb{R}^d} = \lim_{k \rightarrow \infty} (\mathbf{y}_k - \mathbf{x}) / \|\mathbf{y}_k - \mathbf{x}\|_{\mathbb{R}^d}$. From the fact that \mathbf{x} is a local minimizer, there exists a neighborhood B of \mathbf{x} and $\theta \in (0, 1)$ such that

$$\begin{aligned} f_0(\mathbf{y}_k) - f_0(\mathbf{x}) &= \mathcal{L}(\mathbf{y}_k, \boldsymbol{\lambda}) - \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) \\ &= \frac{1}{2} (\mathbf{y}_k - \mathbf{x}) \cdot \{\mathbf{H}_{\mathcal{L}}(\mathbf{x} + \theta(\mathbf{y}_k - \mathbf{x}), \boldsymbol{\lambda}) (\mathbf{y}_k - \mathbf{x})\} \geq 0 \end{aligned} \quad (2.7.33)$$

for arbitrary $\mathbf{y}_k \in B$. Multiplying both sides by $2t / \|\mathbf{y}_k - \mathbf{x}\|_{\mathbb{R}^d}^2$ and taking $k \rightarrow \infty$ yields the result. \square

2.7.4 Sufficient Conditions for Local Minimizers

The following sufficient conditions for yielding local minimizers Problem 2.1.2 can be obtained. $T_{\bar{S}}(\mathbf{x})$ denotes the linearized admissible direction set of \bar{S} defined in Eq. (2.7.32) at \mathbf{x} .

Theorem 2.7.8 (2nd-Order Sufficient Conditions for Local Minimizers) *Let f_0, \dots, f_m be elements of $C^2(X; \mathbb{R})$ in Problem 2.1.2. Given $\mathbf{x} \in X$, let \mathbf{g}_i be linearly independent with respect to $i \in I_A(\mathbf{x})$ and satisfy the first-order constraint*

qualification. If there exists $(\mathbf{x}, \boldsymbol{\lambda})$ satisfying the KKT conditions and if

$$\mathbf{y} \cdot (\mathbf{H}_{\mathcal{L}}(\mathbf{x}, \boldsymbol{\lambda}) \mathbf{y}) > 0$$

for arbitrary $\mathbf{y} \in T_{\bar{S}}(\mathbf{x})$, then \mathbf{x} is a local minimizer. \square

Proof Eq. (2.7.33) is obtained when $(\mathbf{x}, \boldsymbol{\lambda})$ satisfies the KKT conditions. Since $\mathbf{H}_{\mathcal{L}}(\mathbf{x}, \boldsymbol{\lambda})$ is positive definite, \mathbf{x} is a local minimizer. \square

2.7.5 Sufficient Conditions for Global Minimizers Using the KKT Conditions

Sufficient conditions for global minimizers can be obtained when Problem 2.1.2 is a convex optimization problem.

Theorem 2.7.9 (1st-Order Sufficient Conditions for Global Minimizers) *In Problem 2.1.2, let f_0, \dots, f_m be convex functions from $C^1(X; \mathbb{R})$. Given $\mathbf{x} \in X$, assume that g_i is linearly independent with respect to $i \in I_A(\mathbf{x})$, that the linear constraint qualification is satisfied, and that $(\mathbf{x}, \boldsymbol{\lambda})$ satisfies the KKT conditions. If \mathbf{x} satisfies these conditions, then it is a global minimizer.* \square

Proof Fix $\lambda_1, \dots, \lambda_m$ satisfying the KKT conditions and let

$$\mathcal{L}(\mathbf{y}) = f_0(\mathbf{y}) + \sum_{i \in \{1, \dots, m\}} \lambda_i f_i(\mathbf{y}).$$

By the KKT conditions, it follows that $\partial_X \mathcal{L}(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^d}$. Since \mathcal{L} is convex, Corollary 2.7.3 ensures that \mathbf{x} minimizes \mathcal{L} . In other words, the following inequality holds for arbitrary $\mathbf{y} \in S$:

$$f_0(\mathbf{x}) + \sum_{i \in \{1, \dots, m\}} \lambda_i f_i(\mathbf{x}) \leq f_0(\mathbf{y}) + \sum_{i \in \{1, \dots, m\}} \lambda_i f_i(\mathbf{y}).$$

The KKT conditions also imply that $\lambda_i f_i(\mathbf{x}) = 0$ and $\lambda_i \geq 0$ for each $i \in \{1, \dots, m\}$. Therefore

$$f_0(\mathbf{x}) \leq f_0(\mathbf{y})$$

holds for arbitrary $\mathbf{y} \in S$. \square

Theorem 2.7.9 can be used to obtain the following sufficient conditions for showing the existence of global minimizers of optimization problems under equality constraints (Problem 2.6.1).

Corollary 2.7.10 (1st-Order Sufficient Conditions for Global Minimizers) Assume that f_0 is a convex function from $C^1(X; \mathbb{R})$, that $\mathbf{h} \in C^1(X; \mathbb{R}^n)$ is a linear function, and that $\partial_X h_1(\mathbf{x}), \dots, \partial_X h_n(\mathbf{x})$ are linearly independent at $\mathbf{x} \in X$ and satisfy the linear constraint qualifications. If $(\mathbf{x}, \boldsymbol{\lambda})$ is a solution of Problem 2.6.5, then it yields the minimum in Problem 2.6.1. \square

Proof The equality constraint is equivalent to the following two inequalities:

$$\mathbf{h}(\mathbf{x}) \leq \mathbf{0}_{\mathbb{R}^n}, \quad -\mathbf{h}(\mathbf{x}) \leq \mathbf{0}_{\mathbb{R}^n}.$$

Let the Lagrange multipliers with respect to these be $\boldsymbol{\lambda}_+ = (\lambda_{+1}, \dots, \lambda_{+n})^\top \in \mathbb{R}^n$, and $\boldsymbol{\lambda}_- = (\lambda_{-1}, \dots, \lambda_{-n})^\top \in \mathbb{R}^n$. Since h_1, \dots, h_n are linear functions, they are also convex. This follows from the fact that

$$\partial_X h_i(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}) = h_i(\mathbf{y}) - h_i(\mathbf{x})$$

for arbitrary $\mathbf{x} \in X$ and $\mathbf{y} \in X$ (Theorem 2.4.4). Hence, the optimization is convex optimization, including an inequality constraint. The KKT conditions can then be written as

$$\begin{aligned} \mathcal{L}_{\mathbf{x}}(\mathbf{x}, \boldsymbol{\lambda}_+, \boldsymbol{\lambda}_-) &= \mathbf{g}_0(\mathbf{x}) + \partial_X \mathbf{h}^\top(\mathbf{x})(\boldsymbol{\lambda}_+ - \boldsymbol{\lambda}_-) = \mathbf{0}_{\mathbb{R}^d}, \\ \mathcal{L}_{\lambda_{+i}}(\mathbf{x}, \boldsymbol{\lambda}_+, \boldsymbol{\lambda}_-) &= h_i(\mathbf{x}) \leq 0 \quad \text{for } i \in \{1, \dots, n\}, \\ \mathcal{L}_{\lambda_{-i}}(\mathbf{x}, \boldsymbol{\lambda}_+, \boldsymbol{\lambda}_-) &= -h_i(\mathbf{x}) \leq 0 \quad \text{for } i \in \{1, \dots, n\}, \\ \lambda_{+i} h_i(\mathbf{x}) &= 0 \quad \text{for } i \in \{1, \dots, n\}, \quad \lambda_{-i} h_i(\mathbf{x}) = 0 \quad \text{for } i \in \{1, \dots, n\}, \\ \lambda_{+i} &\geq 0 \quad \text{for } i \in \{1, \dots, n\}, \quad \lambda_{-i} \geq 0 \quad \text{for } i \in \{1, \dots, n\}. \end{aligned}$$

Writing $\boldsymbol{\lambda}_+ - \boldsymbol{\lambda}_- = \boldsymbol{\lambda}$, we see that $(\mathbf{x}, \boldsymbol{\lambda})$ is equivalent to the solution of Problem 2.6.5. \square

2.7.6 Example of an Optimization Problem Under an Inequality Constraint

Let us now consider the KKT conditions in relation to the spring combination problem.

Exercise 2.7.11 (Spring Combination Problem) Consider Exercise 2.6.8 and find the global minimizer when the spring combination conditions are changed to inequalities. In other words, find \mathbf{u} satisfying the following minimization problem:

$$\min_{\mathbf{u} \in \mathbb{R}^2} \left\{ f_0(\mathbf{u}) = \frac{1}{2}k_1u_1^2 + \frac{1}{2}k_2u_2^2 \mid f_1(\mathbf{u}) = a - (u_1 + u_2) \leq 0 \right\}.$$

Also find \mathbf{u} satisfying

$$\min_{\mathbf{u} \in \mathbb{R}^2} \left\{ f_0(\mathbf{u}) = \frac{1}{2}k_1u_1^2 + \frac{1}{2}k_2u_2^2 \mid f_1(\mathbf{u}) = (u_1 + u_2) - a \leq 0 \right\}. \quad \square$$

Answer When $f_1 = a - (u_1 + u_2) \leq 0$, if we let $\lambda \in \mathbb{R}$ be the Lagrange multiplier, then the Lagrange function becomes

$$\mathcal{L}(\mathbf{u}, \lambda) = \frac{1}{2}k_1u_1^2 + \frac{1}{2}k_2u_2^2 + \lambda(a - u_1 - u_2).$$

The stationary conditions of $\mathcal{L}(\mathbf{u}, \lambda)$ are the same as the results from Exercise 2.6.8, and when $k_1 > 0, k_2 > 0$ and $a > 0$ we obtain

$$u_1 = \frac{k_2}{k_1 + k_2}a > 0, \quad u_2 = \frac{k_1}{k_1 + k_2}a > 0, \quad \lambda = \frac{k_1k_2}{k_1 + k_2}a > 0.$$

This result satisfies the KKT conditions. As investigated in Exercise 2.6.8, this problem is also a convex optimization problem and therefore, by Theorem 2.7.9, \mathbf{u} yields the minimum.

On the other hand, when $f_1 = (u_1 + u_2) - a \leq 0$, the Lagrange function becomes

$$\mathcal{L}(\mathbf{u}, \lambda) = \frac{1}{2}k_1u_1^2 + \frac{1}{2}k_2u_2^2 + \lambda(u_1 + u_2 - a).$$

When $k_1 > 0, k_2 > 0$ and $a > 0$, the stationary conditions for $\mathcal{L}(\mathbf{u}, \lambda)$ are

$$u_1 = \frac{k_2}{k_1 + k_2}a > 0, \quad u_2 = \frac{k_1}{k_1 + k_2}a > 0, \quad \lambda = -\frac{k_1k_2}{k_1 + k_2}a < 0.$$

Since $\lambda < 0$, this result does not satisfy the KKT conditions. Therefore, the coupled constraints can be viewed as inactive and we can set $\lambda = 0$. The problem can then be rewritten as

$$\min_{\mathbf{u} \in \mathbb{R}^2} \left\{ f_0(\mathbf{u}) = \frac{1}{2}k_1u_1^2 + \frac{1}{2}k_2u_2^2 \right\}.$$

Here, since

$$\mathbf{g}_0(\mathbf{u}) = \begin{pmatrix} k_1 & 0 \\ 0 & k_2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

we obtain $\mathbf{u} = \mathbf{0}_{\mathbb{R}^2}$. \square

2.7.7 Considerations Relating to the Solutions of Optimization Problems Under Inequality Constraints

The results of this section lead to the following observations regarding the solution of optimization problems under inequality constraints (Problem 2.1.2):

- (1) By Theorem 2.7.5, the solution $(\mathbf{x}, \boldsymbol{\lambda})$ of the Lagrange multiplier method (Problem 2.7.6) satisfies a necessary condition for attaining a local minimum. Such \mathbf{x} are candidates for local minimizers.
- (2) When $(\mathbf{x}, \boldsymbol{\lambda})$ is the solution from the Lagrange multiplier method (Problem 2.7.6), if the Hesse matrix $\partial_X \partial_X^\top \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{H}_{\mathcal{L}}(\mathbf{x}, \boldsymbol{\lambda})$ of the Lagrange function with respect to \mathbf{x} satisfies $\mathbf{y} \cdot (\mathbf{H}_{\mathcal{L}}(\mathbf{x}, \boldsymbol{\lambda}) \mathbf{y}) > 0$ for arbitrary $\mathbf{y} \in T_{\mathcal{S}}(\mathbf{x})$, then Theorem 2.7.8 implies that \mathbf{x} is a local minimizer.
- (3) When an optimization problem under an inequality constraint (Problem 2.1.2) is convex, Theorem 2.7.9 implies that the solution \mathbf{x} from the Lagrange multiplier method yields the global minimum.

2.8 Optimization Problems Under Equality and Inequality Constraints

In optimal design problems, state equations are set using equality constraints, and cost function constraints are set through inequality constraints. The optimization problem defined at the beginning of Sect. 2.1 (Problem 2.1.1) was defined with this in mind. Keeping in mind its correspondence with optimal design problems, we write Problem 2.1.1 in the following way.

Problem 2.8.1 (Optimization Under Equality and Inequality Constraints)
Given $d > n$, let $\Xi = \mathbb{R}^{d-n}$ and $U = \mathbb{R}^n$. When $\mathbf{K} : \Xi \rightarrow \mathbb{R}^{n \times n}$ and $\mathbf{b} : \Xi \rightarrow \mathbb{R}^n$ are given together with $f_0, f_1, \dots, f_m : \Xi \times U \rightarrow \mathbb{R}$, find $(\boldsymbol{\xi}, \mathbf{u})$ satisfying

$$\begin{aligned} \min_{(\boldsymbol{\xi}, \mathbf{u}) \in \Xi \times U} \{ & f_0(\boldsymbol{\xi}, \mathbf{u}) \mid \mathbf{h}(\boldsymbol{\xi}, \mathbf{u}) = -\mathbf{K}(\boldsymbol{\xi})\mathbf{u} + \mathbf{b}(\boldsymbol{\xi}) = \mathbf{0}_{\mathbb{R}^n}, \\ & f_1(\boldsymbol{\xi}, \mathbf{u}) \leq 0, \dots, f_m(\boldsymbol{\xi}, \mathbf{u}) \leq 0 \}. \end{aligned}$$

□

Consider Problem 2.8.1 and let the set of $(\boldsymbol{\xi}, \mathbf{u})$ satisfying the equality constraints be given by

$$V = \{(\boldsymbol{\xi}, \mathbf{u}) \in \Xi \times U \mid \mathbf{h}(\boldsymbol{\xi}, \mathbf{u}) = -\mathbf{K}(\boldsymbol{\xi})\mathbf{u} + \mathbf{b}(\boldsymbol{\xi}) = \mathbf{0}_{\mathbb{R}^n}\}.$$

For $i \in \{0, 1, \dots, m\}$ let

$$\tilde{f}_i(\boldsymbol{\xi}) = \{f_i(\boldsymbol{\xi}, \mathbf{u}) \mid (\boldsymbol{\xi}, \mathbf{u}) \in V\}. \quad (2.8.1)$$

Then the derivative of $\tilde{f}_i(\xi)$ with respect to ξ can be obtained:

$$\tilde{\mathbf{g}}_i = \frac{\partial f_i}{\partial \xi} + \left(\frac{\partial \mathbf{b}}{\partial \xi^\top} - \frac{\partial \mathbf{K}}{\partial \xi^\top} \mathbf{u} \right)^\top \mathbf{v}_i \in \mathbb{R}^{d-n}, \quad (2.8.2)$$

The above is arrived at in the same manner as Eq. (2.6.25) from Sect. 2.6.5. Here, $\left(\frac{\partial \mathbf{K}}{\partial \xi^\top} \right) \mathbf{u}$ is defined by Eq. (2.6.21) and $\mathbf{v}_i \in U$ is the solution of the equivalent adjoint problem to Eq. (2.6.23):

$$\mathbf{K}^\top \mathbf{v}_i = \frac{\partial f_i}{\partial \mathbf{u}}. \quad (2.8.3)$$

The functions $\tilde{\mathbf{g}}_0, \dots, \tilde{\mathbf{g}}_m$ obtained in this way are the derivatives with respect to ξ when the cost functions are taken to be $\tilde{f}_0(\xi), \dots, \tilde{f}_m(\xi)$. These facts allow one to rewrite Problem 2.8.1 as follows.

Problem 2.8.2 (Optimization Under Inequality Constraints) Let $\Xi = \mathbb{R}^{d-n}$ and $\tilde{f}_0, \tilde{f}_1, \dots, \tilde{f}_m : \Xi \rightarrow \mathbb{R}$ be given. Find ξ satisfying

$$\min_{\xi \in \Xi} \left\{ \tilde{f}_0(\xi) \mid \tilde{f}_1(\xi) \leq 0, \dots, \tilde{f}_m(\xi) \leq 0 \right\}. \quad \square$$

Let the Lagrange function with respect to Problem 2.8.2 be

$$\tilde{\mathcal{L}}(\xi, \lambda) = \tilde{f}_0(\xi) + \sum_{i \in \{1, \dots, m\}} \lambda_i \tilde{f}_i(\xi). \quad (2.8.4)$$

Then the KKT conditions with respect to Problem 2.8.2 become

$$\tilde{\mathcal{L}}_\xi(\xi, \lambda) = \tilde{\mathbf{g}}_0 + \sum_{i \in \{1, \dots, m\}} \lambda_i \tilde{\mathbf{g}}_i = \mathbf{0}_{\mathbb{R}^{d-n}}, \quad (2.8.5)$$

$$\tilde{\mathcal{L}}_{\lambda_i}(\xi, \lambda) = \tilde{f}_i(\xi) \leq 0 \quad \text{for } i \in \{1, \dots, m\}, \quad (2.8.6)$$

$$\lambda_i \tilde{f}_i(\xi) = 0 \quad \text{for } i \in \{1, \dots, m\}, \quad (2.8.7)$$

$$\lambda_i \geq 0 \quad \text{for } i \in \{1, \dots, m\}. \quad (2.8.8)$$

Therefore, the Lagrange multiplier method for seeking candidates for local minimizers of Problem 2.8.2 can be expressed as follows.

Problem 2.8.3 (Lagrange Multiplier Method with Inequality Constraints) Let $\tilde{\mathcal{L}}(\xi, \lambda)$ be given by Eq. (2.8.4) with respect to Problem 2.8.2. Find (ξ, λ) satisfying the KKT conditions (Eqs. (2.8.5)–(2.8.8)). \square

Therefore the results of Theorems 2.7.5–2.7.8 can be obtained with respect to the solution (ξ, λ) of Problem 2.8.3. We will postpone our discussion of what these

results say about the solution of optimization problems under equality and inequality constraints (Problem 2.8.1) until Sect. 2.8.2.

2.8.1 The Lagrange Multiplier Method for Optimization Problems Under Equality and Inequality Constraints

The conditions satisfied by the minimizer of Problem 2.8.1 are as described above. Let us now define the Lagrange function for Problem 2.8.1 and consider how it can be related to the Lagrange multiplier method for Problem 2.8.3. We remark that the content shown here was also shown in Chap. 1. The concepts are the same as those used in deriving derivatives of cost functionals with respect to design variables in the optimal design problems of Chap. 7 and beyond. The purpose of this section is to clarify the relationship of such concepts with the content of Chap. 2.

Let the Lagrange function with respect to Problem 2.8.1 be given by

$$\mathcal{L}(\xi, u, v_0, \dots, v_m, \lambda) = \mathcal{L}_0(\xi, u, v_0) + \sum_{i \in \{1, \dots, m\}} \lambda_i \mathcal{L}_i(\xi, u, v_i), \quad (2.8.9)$$

where $\lambda = (\lambda_1, \dots, \lambda_m)^\top \in \mathbb{R}^m$ denotes the Lagrange multiplier with respect to $f_1 \leq 0, \dots, f_m \leq 0$. Also, let

$$\mathcal{L}_i(\xi, u, v_i) = f_i(\xi, u) + \mathcal{L}_S(\xi, u, v_i) \quad (2.8.10)$$

denote the Lagrange function with respect to $f_i(\xi, u)$. Moreover, let the Lagrange function with respect to the equality constraint be given by

$$\mathcal{L}_S(\xi, u, v_i) = -v_i \cdot (K(\xi)u - b(\xi)), \quad (2.8.11)$$

and v_0, \dots, v_m be Lagrange multipliers (adjoint variables) defined for f_0, \dots, f_m , respectively.

The function \tilde{g}_i from Eq. (2.8.2) is obtained as follows. The derivative (total differential) of \mathcal{L}_i with respect to an arbitrary variation $(\eta, \hat{u}, \hat{v}_i) \in \Xi \times U \times U$ of (ξ, u, v_i) is given by

$$\begin{aligned} \mathcal{L}'_i(\xi, u, v_i)[\eta, \hat{u}, \hat{v}_i] &= \mathcal{L}_{i\xi}(\xi, u, v_i)[\eta] + \mathcal{L}_{iu}(\xi, u, v_i)[\hat{u}] \\ &\quad + \mathcal{L}_{iv_i}(\xi, u, v_i)[\hat{v}_i]. \end{aligned} \quad (2.8.12)$$

The third term on the right-hand side of Eq. (2.8.12) becomes

$$\mathcal{L}_{iv_i}(\xi, u, v_i)[\hat{v}_i] = -\hat{v}_i \cdot (K(\xi)u - b(\xi)) = \mathcal{L}_S(\xi, u, \hat{v}_i). \quad (2.8.13)$$

Equation (2.8.13) takes the value zero when \mathbf{u} satisfies the equality constraint. The second term on the right-hand side of Eq. (2.8.12) is expressed as

$$\mathcal{L}_{iu}(\xi, \mathbf{u}, \mathbf{v}_i)[\hat{\mathbf{u}}] = -\hat{\mathbf{u}} \cdot \left(\mathbf{K}^\top(\xi) \mathbf{v}_i - \frac{\partial f_i}{\partial \mathbf{u}} \right). \quad (2.8.14)$$

Equation (2.8.14) also takes the value zero when \mathbf{v}_i satisfies Eq. (2.8.3). The first term on the right-hand side of Eq. (2.8.12) is given by

$$\mathcal{L}_{i\xi}(\xi, \mathbf{u}, \mathbf{v}_i)[\eta] = \left(\frac{\partial f_i}{\partial \xi} + \left(\frac{\partial \mathbf{b}}{\partial \xi^\top} - \frac{\partial \mathbf{K}}{\partial \xi^\top} \mathbf{u} \right)^\top \mathbf{v}_i \right) \cdot \eta. \quad (2.8.15)$$

The above results show that placing the equality constraint on \mathbf{u} is equivalent to the condition that $\mathcal{L}_{iv_i}(\xi, \mathbf{u}, \mathbf{v}_i)[\hat{\mathbf{v}}_i] = 0$ for all $\hat{\mathbf{v}}_i \in U$, and that placing an adjoint equation on \mathbf{v}_i is equivalent to $\mathcal{L}_{iu}(\xi, \mathbf{u}, \mathbf{v}_i)[\hat{\mathbf{u}}] = 0$ for all $\hat{\mathbf{u}} \in U$. These results also show that \mathbf{u} and \mathbf{v}_i can be used to obtain $\tilde{f}'_i(\xi)[\eta]$ (the derivative of \tilde{f}_i) from $\mathcal{L}_{i\xi}(\xi, \mathbf{u}, \mathbf{v}_i)[\eta] = \tilde{g}_i \cdot \eta$.

2.8.2 Considerations Regarding Optimization Problems Under Equality and Inequality Constraints

Based on the results obtained thus far, the following can be said with respect to the solution of optimization problems (Problem 2.8.1) under equality and inequality constraints:

- (1) Let $\tilde{f}_0, \dots, \tilde{f}_m$ be given by Eq. (2.8.1). Then an optimization problem under equality and inequality constraints (Problem 2.8.1) can be rewritten as an optimization problem under an inequality constraint (Problem 2.8.2).
- (2) Let $i \in \{0, 1, \dots, m\}$ and the Lagrange function of Problem 2.8.1 be given by \mathcal{L}_i from Eq. (2.8.10). Then $\tilde{f}'_i(\xi)[\eta]$ (the derivative of \tilde{f}_i) can be obtained from $\mathcal{L}_{i\xi}(\xi, \mathbf{u}, \mathbf{v}_i)[\eta] = \tilde{g}_i \cdot \eta$ by using \mathbf{u} and \mathbf{v}_i which satisfy the equality constraint and the adjoint equation, respectively (or, by using $\mathcal{L}_{iu}(\xi, \mathbf{u}, \mathbf{v}_i)[\hat{\mathbf{u}}] = 0$ with respect arbitrary $\hat{\mathbf{u}} \in U$).
- (3) Let (ξ, λ) denote the solution of an optimization problem under an inequality constraint (Problem 2.8.2) which has been obtained via the Lagrange multiplier method (Problem 2.8.3) (details of the methodology are given in Chap. 3). When the Hesse matrix $\partial_\xi \partial_\xi^\top \tilde{\mathcal{L}}(\xi, \lambda) = \mathbf{H}_{\tilde{\mathcal{L}}}(\xi, \lambda)$ with respect to ξ of the Lagrange function $\tilde{\mathcal{L}}(\xi, \lambda)$ satisfies

$$\eta \cdot (\mathbf{H}_{\tilde{\mathcal{L}}}(\xi, \lambda) \eta) > 0$$

for all variations η belonging to

$$T_{\tilde{S}}(\xi) = \{\eta \in \Xi \mid \tilde{g}_i(\xi) \cdot \eta = 0 \text{ for all } i \in I_A(\xi)\},$$

then Theorem 2.7.8 implies that ξ is a local minimizer.

- (4) When optimization problems include an inequality constraint (Problem 2.8.2) and are convex, Theorem 2.7.9 implies that the Lagrange multiplier method's solution ξ and the function u satisfying $h(\xi, u) = \mathbf{0}_{\mathbb{R}^n}$ are global minimizers.

A one-dimensional linear elastic body and a one-dimensional steady Stokes flow field were used as examples of an optimization problem under inequality constraints in Chap. 1. These problems are convex optimization problems. Hence, if a satisfying the KKT conditions can be found, then it can be deemed to minimize the problems.

2.9 Duality Theorem

The KKT conditions used in Sect. 2.7 and Sect. 2.8 required that f_0, \dots, f_m be first-order differentiable. The duality theorem (shown next) allows one to replace first-order differentiability with convexity. Since this theorem is not directly used in this book the result will be stated without proof.

Let us define the constraint qualification as follows.

Definition 2.9.1 (Slater Constraint Qualification) In Problem 2.1.2, if there exists $y \in S$ such that $f(y) < \mathbf{0}_{\mathbb{R}^m}$, then we say that the Slater constraint qualification is satisfied. \square

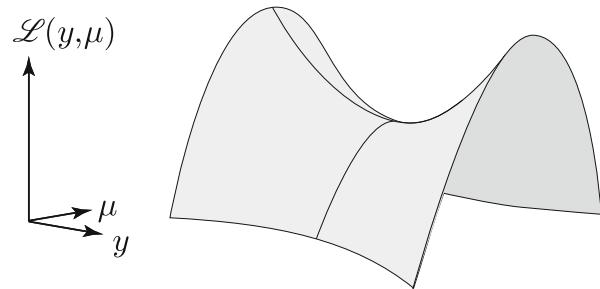
The duality theorem is expressed as follows [50, 51, 59].

Theorem 2.9.2 (Duality Theorem) Suppose that Problem 2.1.2 is a convex optimization problem and the Slater constraint qualification is satisfied. Let $\mathcal{L}(x, \lambda)$ be given by Eq. (2.7.26). Here, the necessary and sufficient condition for $x \in X$ to yield the minimum is for there to exist $\lambda \geq \mathbf{0}_{\mathbb{R}^m}$ such that the following holds for arbitrary $y \in X$ and $\mu \geq \mathbf{0}_{\mathbb{R}^m}$:

$$\mathcal{L}(x, \mu) \leq \mathcal{L}(x, \lambda) \leq \mathcal{L}(y, \lambda). \quad (2.9.1)$$

\square

A pair (x, λ) which satisfies Eq. (2.9.1) describes a saddle point such as the one shown in Fig. 2.22. For this reason, the duality theorem is also referred to as the saddle point theorem.

Fig. 2.22 Saddle point

2.9.1 Examples of the Duality Theorem

Let us make use of the duality theorem with respect to the following combined spring problem.

Exercise 2.9.3 (Spring Combination Problem) Consider Exercise 2.7.11 and show that the \mathbf{u} satisfying

$$\min_{\mathbf{u} \in \mathbb{R}^2} \left\{ f_0(\mathbf{u}) = \frac{1}{2}k_1u_1^2 + \frac{1}{2}k_2u_2^2 \mid f_1(\mathbf{u}) = a - (u_1 + u_2) \leq 0 \right\}$$

is a saddle point of the Lagrange function. \square

Answer Let \mathbf{u} be the minimizer for this problem and $\mathbf{v} \in \mathbb{R}^2$ be arbitrary. Since $f_0(\mathbf{v})$ and $f_1(\mathbf{v})$ are convex functions, the optimization problem is convex. Moreover, from the fact that $f_1(\mathbf{v}) < 0$ when $(v_1, v_2) = (a/4, a/4)$, the Slater constraint qualification is satisfied. The Lagrange function for this problem is defined by

$$\mathcal{L}(\mathbf{v}, \mu) = f_0(\mathbf{v}) + \mu f_1(\mathbf{v}) = \frac{1}{2}k_1v_1^2 + \frac{1}{2}k_2v_2^2 + \mu(a - v_1 - v_2),$$

where $\mu \in \mathbb{R}$ is a Lagrange multiplier with respect to $f_1 \leq 0$. For $\mu > 0$, we have

$$\begin{aligned} \mathcal{L}(\mathbf{u}, \mu) &= \inf_{\mathbf{v} \in \mathbb{R}^2} \mathcal{L}(\mathbf{v}, \mu) = \mathcal{L}\left(\frac{\mu}{k_1}, \frac{\mu}{k_2}, \mu\right) = \mathcal{L}_S(\mu) \\ &= \frac{1}{2}\frac{\mu^2}{k_1} + \frac{1}{2}\frac{\mu^2}{k_2} + \mu\left(a - \frac{\mu}{k_1} - \frac{\mu}{k_2}\right) = -\frac{1}{2}\left(\frac{1}{k_1} + \frac{1}{k_2}\right)\mu^2 + a\mu. \end{aligned}$$

We let $\tilde{\mathcal{L}}(\mu) = \mathcal{L}(\mathbf{u}(\mu), \mu)$ and set λ equal to the μ satisfying

$$\frac{d\tilde{\mathcal{L}}}{d\mu} = -\left(\frac{1}{k_1} + \frac{1}{k_2}\right)\mu + a = 0.$$

We also have

$$\frac{d^2 \tilde{\mathcal{L}}}{d\mu^2} = -\left(\frac{1}{k_1} + \frac{1}{k_2}\right) < 0.$$

Hence,

$$\mathcal{L}(\mathbf{u}, \mu) \leq \mathcal{L}(\mathbf{u}, \lambda).$$

On the other hand, it is easy to see that $\mathcal{L}(\mathbf{u}, \lambda) \leq \mathcal{L}(\mathbf{v}, \lambda)$. \square

Let us now visually confirm that the minimum in Exercise 2.9.3 is a saddle point of the Lagrange function. The variable in this problem was $(u_1, u_2, \lambda)^\top \in X = \mathbb{R}^3$. The current setting is difficult to illustrate. Hence we take $u_2 = 0$ (or, equivalently, $k_2 \rightarrow \infty$) and let $k_1 = 1$ and $a = 1$. Then

$$\mathcal{L}(u_1, \lambda) = \frac{1}{2}u_1^2 + \lambda(1 - u_1)$$

and the saddle point becomes $(u_1, \lambda) = (1, 1)$. Figure 2.23 shows the situation, from which we confirm the saddle point.

Here, we remark that λ refers to an internal force and that $-\mathcal{L}(\mathbf{u}, \lambda)$ is a complementary energy. Minimization of complementary energy is used in engineering when seeking internal forces from a given displacement.

Let us end this chapter with an application of the duality theorem with respect to an optimal design problems such as was treated in Chap. 1. As our observations in Sect. 1.1 and Sect. 2.2 showed, the optimal design problems considered in this book have been convex optimization problems. The duality theorem is thus applicable and the Lagrange function should form a saddle point with respect to the design variable

Fig. 2.23 The saddle point in Exercise 2.9.3

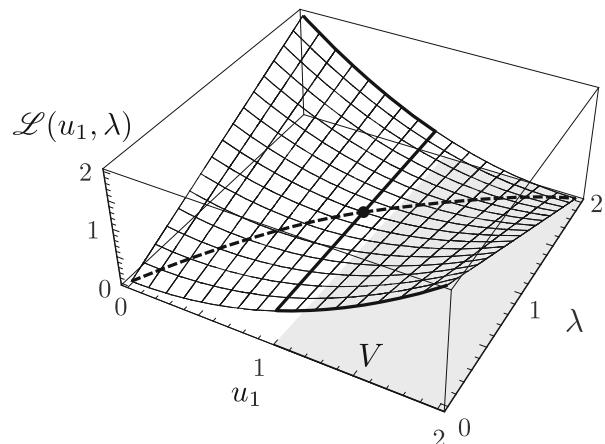
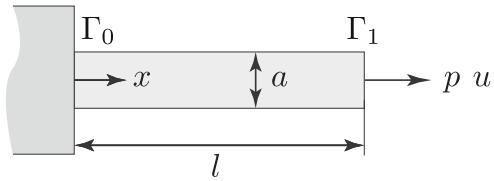


Fig. 2.24 One-dimensional elastic body with one cross-section



and Lagrange multiplier at the minimizer of the optimal design problem. Let us take a look at this using a diagram.

In order to aid the illustration, we limit the number of variables to two. If one of the variables is a Lagrange multiplier, the number of design variables must be limited to one. Therefore, let us consider a one-dimensional linear elastic body with a single cross-sectional area, such as the one shown in Fig. 2.24.

Exercise 2.9.4 (Mean Compliance Minimization Problem) Suppose that $e_Y = 1$, $l = 1$, $c_1 = 1$ and $p = 1$. Find (a, u) satisfying

$$\min_{(a,u) \in \mathbb{R}^2} \left\{ f_0(u) = pu \mid f_1(a) = la - c_1 \leq 0, \frac{e_Y}{l} au = p \right\}.$$

Also graph the Lagrange function at this point on a diagram. \square

Answer We have $\tilde{f}_0(a) = f_0(u(a)) = f_0(1/a) = 1/a$. The functions $\tilde{f}_0(a)$ and $f_1(a)$ are convex, and therefore

$$\min_{a \in \mathbb{R}} \left\{ \tilde{f}_0(a) \mid f_1(a) \leq 0 \right\}$$

is a convex optimization problem. The Slater constraint qualification is clearly satisfied. The Lagrange function for this problem is given by

$$\mathcal{L}(a, \lambda) = \tilde{f}_0(a) + \lambda f_1(a) = \frac{1}{a} + \lambda(a - 1),$$

where $\lambda \in \mathbb{R}$ is a Lagrange multiplier with respect to $f_1(a) \leq 0$. We have

$$\mathcal{L}_a = -\frac{1}{a^2} + \lambda = 0, \quad \mathcal{L}_\lambda = a - 1 = 0,$$

and it follows that $(a, \lambda) = (1, 1)$ is a stationary point of $\mathcal{L}(a, \lambda)$. Figure 2.25 shows \mathcal{L} in a neighborhood of this point, from which we can confirm the existence of the saddle point. \square

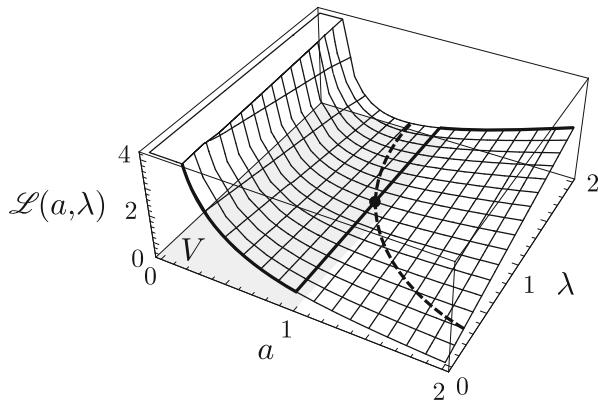


Fig. 2.25 Saddle point of Exercise 2.9.4

2.10 Summary

Chapter 2 examined the theory of optimization problems in finite-dimensional vector spaces. The key points are as follows.

- (1) Optimization problems are generally defined as finding an element at which a cost function attains its minimum value amongst a set of design variables. On the other hand, as we saw in Chap. 1, state and design variables are defined in optimal design problems, and cost functions are defined as functions of the design and state variables. In this case, state variables are uniquely determined via state equations. Here, in order to fit the optimal design problem into the framework of a general optimization problem, a state variable can be included in the design variable, and state equations should be viewed as equality constraints (Sect. 2.1).
- (2) The gradient of a cost function is 0 at local minimizers of unconstrained optimization problems (Theorem 2.5.2). Moreover, if the Hesse matrix is positive definite at the stationary point (gradient 0) of the cost function, then it is a local minimizer (Theorem 2.5.5). Furthermore, if the cost function is convex, the local minimizer yields the global minimum (Theorem 2.5.6 or Corollary 2.7.3).
- (3) At a local minimum of an optimization problem under equality constraints, the Lagrange function is stationary (Theorem 2.6.4). Also, when the Hesse matrix of the Lagrange function is positive definite with respect to variation of a variable satisfying an equality constraint at a stationary point of the Lagrange function, it follows that the stationary point is a local minimizer (Theorem 2.6.7). Furthermore, if the optimization problem is convex, the stationary point of the Lagrange function is a global minimizer (Corollary 2.7.10).
- (4) The KKT conditions hold at local minimizers of optimization problems under inequality constraints (Theorem 2.7.5). If the Hesse matrix of the Lagrange

function with respect to arbitrary variation of a variable satisfying an inequality constraint is positive definite at the point where the KKT condition is satisfied, then that point is a local minimizer (Theorem 2.7.8). Furthermore, if the optimization problem is convex, a point satisfying the KKT condition yields the global minimum (Theorem 2.7.9).

- (5) At local minimizers of optimization problems under equality and inequality constraints, KKT conditions are established from the derivative of the cost function with respect to the independent variation of an unconstrained variable while an equality constraint is satisfied (Eqs. (2.8.5)–(2.8.8)).
- (6) Minimizers of convex optimization problems which include inequality constraints form saddle points of their Lagrange functions (Theorem 2.9.2).

The literature of optimization theory is vast. In addition to the references cited in this chapter we also refer to [22, 40, 79, 106, 162, 179].

2.11 Practice Problems

2.1 In Definition 2.4.5, if A is positive definite, show that α equals the minimum value of the eigenvalues of A . Also show that if A is negative definite, that $-\alpha$ equals the maximum value of the eigenvalues of A . (Hint: Refer to Theorem A.2.1.)

2.2 Let $f : \mathbb{R}^2 \mapsto \mathbb{R}$ be the function

$$f(x_1, x_2) = \frac{1}{2} (ax_1^2 + 2bx_1x_2 + cx_2^2) + dx_1 + ex_2,$$

where $a, b, c, d \in \mathbb{R}$ are constants. Derive necessary conditions for f to attain a minimum value. Also show that a sufficient condition is $a > 0$ and $ac - b^2 > 0$. The Sylvester criterion (Theorem A.2.2) may of course be used.

2.3 Amongst the set of rectangles whose perimeter is less than a given value, show that the one with the largest area is a square.

- Let the length of the sides of the rectangle be expressed by $x = (x_1, x_2)^\top \in \mathbb{R}^2$. Construct the problem by defining a positive real constant constraining the perimeter length to be c_1 .
- Define the Lagrange function and find the KKT conditions.
- Show that if a solution satisfies the KKT conditions, then it is a global minimizer. (Hint: Consider whether or not this is a convex optimization problem. If it is not, show that the problem in which the cost function is recreated using functions (denoted by \tilde{f}_0 here) with respect to the set of constrained variables is a convex optimization problem.)

Chapter 3

Basics of Mathematical Programming



In Chap. 2, we discussed the conditions satisfied by a local minimum point (the required conditions of a local minimum point) and the conditions which guarantee it to be a minimum point (sufficient conditions for a minimum point) under a finite-dimensional vector space setting. No detailed explanation, however, was provided regarding the method (solution) for finding the local minimum point. In this chapter, we would like to address this ensuing matter. The computational formulation associated to such a problem is called an optimization problem or a mathematical programming problem, and active research is being conducted in the academic field referred to as operations research (OR). Here, we will consider algorithms while showing results that are theoretically obtained or ways to deal with the solution of optimization problems. Much of the content covered here is also valid for abstract optimal design problems in Chap. 7. In fact, in Chap. 7 we will see how the same algorithms can be adapted for function spaces.

In this chapter, we assume that the cost functions (objective and constraint functions) have computable gradients and Hessians. However, the real difficulty from the computational point of view is how to evaluate them. We want the reader bears this in mind.

3.1 Problem Setting

Optimal design problems, as seen in Chap. 1, were viewed as optimization problems with equality constraints (state equations) and inequality constraints of the cost functions $f_0(\xi, \mathbf{u}), \dots, f_m(\xi, \mathbf{u})$ defined by the design variable $\xi \in \Xi$ and state variable $\mathbf{u} \in U$. In Chap. 2, such a problem was seen as an optimization problem constructed as $f_0(\xi, \mathbf{u}), \dots, f_m(\xi, \mathbf{u})$ with $\mathbf{x} = (\xi, \mathbf{u}) \in \Xi \times U$ as a design variable.

In this chapter, we shall recall the definitions given in Chap. 1 and let ξ be the design variable with $f_0(\xi, \mathbf{u}(\xi)), \dots, f_m(\xi, \mathbf{u}(\xi))$ denoted by $\tilde{f}_0(\xi), \dots, \tilde{f}_m(\xi)$, respectively. The differential of $\tilde{f}_0(\xi), \dots, \tilde{f}_m(\xi)$ with respect to ξ can be obtained via adjoint variable method as seen in Sect. 2.8. Furthermore, $\tilde{f}_0(\xi), \dots, \tilde{f}_m(\xi)$ are assumed to be non-linear functions. Actually, in the optimal design problem of Chap. 1 (Problem 1.1.4), even if $f_0(\mathbf{u})$ is a linear function with respect to \mathbf{u} , the equality constraint function $\mathbf{h}(\mathbf{a}, \mathbf{u}) = -\mathbf{K}(\mathbf{a})\mathbf{u} + \mathbf{p}$ is non-linear with respect to (\mathbf{a}, \mathbf{u}) , hence, $\tilde{f}_0(\mathbf{a})$ became a non-linear function.

In this chapter, by denoting the design variable $\xi \in \Xi$ as $\mathbf{x} \in X = \mathbb{R}^d$, the non-linear functions $\tilde{f}_0, \dots, \tilde{f}_m$ as f_0, \dots, f_m , and the gradient of these with respect to \mathbf{x} as $\mathbf{g}_0, \dots, \mathbf{g}_m$, respectively, we can consider the following problem which does not include any equality constraints.

Problem 3.1.1 (Non-linear Optimization Problem) Let $X = \mathbb{R}^d$. Given the functions $f_0, \dots, f_m \in C^1(X; \mathbb{R})$, find an element \mathbf{x} which satisfies

$$\min_{\mathbf{x} \in X} \{ f_0(\mathbf{x}) \mid f_1(\mathbf{x}) \leq 0, \dots, f_m(\mathbf{x}) \leq 0 \}. \quad \square$$

The structure of this chapter is as follows. In Sect. 3.2, the definitions relating to convergence and the definition of iterative method, which is a basic way to think about the solutions of non-linear optimization problems, will be presented. Then, from Sects. 3.3 to 3.5, we will look at solutions with respect to unconstrained optimization problems. After that, we will discuss the solutions of optimization problems with inequality constraints (Problem 3.1.1) in Sect. 3.6 and in the rest of the chapter.

3.2 Iterative Method

Given the solutions of non-linear optimization problems, there do not appear to be any methods which allow us to obtain the optimal solution by solving simultaneous linear equations once without any pre-processing. Usually the iterative method shown below is the standard.

Definition 3.2.1 (Iterative Method) A method whereby a non-minimum point $\mathbf{x}_0 \in X$ is chosen with respect to Problem 3.1.1, and seeking

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{y}_g = \mathbf{x}_k + \bar{\epsilon}_g \bar{\mathbf{y}}_g \quad (3.2.1)$$

with respect to $k \in \mathbb{N}$, while determining $\mathbf{y}_g \in X$ is called an iterative method. Here, \mathbf{y}_g is called a search vector and its size $\|\mathbf{y}_g\|_X$, a step size. In contrast, $\bar{\mathbf{y}}_g$ is a vector providing direction only, and in this book, it is distinguished from a search vector by referring to it as the search direction. It is assumed that the size of $\bar{\mathbf{y}}_g$ does

not need to be one. $\bar{\epsilon}_g$ is a positive constant for adjusting its size. Moreover, \mathbf{x}_0 is called the initial point and \mathbf{x}_k , where $k \in \mathbb{N}$, is called a trial point. \square

From this definition, given an algorithm using iterative methods, there is a need to specify the methods for seeking the search direction $\bar{\mathbf{y}}_g$ and a method to appropriately determine the step size $\|\mathbf{y}_g\|_X$. We will look at these methods in Sect. 3.3 onwards. Moreover, aside from this iterative method, there is a known numerical solution to optimization problems called the direct method. The direct method is used as a collective term for methods which allow solutions to be sought via a finite number of steps. This method, however, will not be discussed in this book since it is mainly used to deal with linear optimization problems.

For the purpose of later discussions, a glossary representing the characteristics and qualities of the iterative method will be defined.

Definition 3.2.2 (Global Convergence) An iterative method is said to have global convergence when an initial point is arbitrarily chosen and yet it generates a sequence of iterates that converges to a point for which a necessary condition of optimality holds. \square

Definition 3.2.3 (Convergence Rate) Let \mathbf{x} be a local minimum and $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$ be a sequence of iterates obtained through an iterative method. If there exists an index k_0 and a constant $p \in [1, \infty)$ such that the inequality condition

$$\|\mathbf{x}_{k+1} - \mathbf{x}\|_X \leq \|\mathbf{x}_k - \mathbf{x}\|_X^p$$

holds for each $k \geq k_0$, then p is called the convergence order of the algorithm. Here, when $p = 1$, $r \in (0, 1)$ and when $p > 1$, r is a positive constant. Moreover, when r can be replaced by a number sequence $\{r_k\}_{k \in \mathbb{N}}$ which converges to zero, the algorithm exhibits a super p th order of convergence. \square

3.3 Gradient Method

Let us first examine the gradient method as a procedure to find the search direction $\bar{\mathbf{y}}_g$. Here, let us consider choosing one cost function f_i from $i \in \{0, 1, \dots, m\}$ and obtain the direction in which $\bar{\mathbf{y}}_{gi}$ descends. Such a $\bar{\mathbf{y}}_{gi}$ will be referred to as the descent direction of f_i .

If in Problem 3.1.1, the minimum point is a point within the admissible set (all inequality constraints become inactive), the descent direction $\bar{\mathbf{y}}_{g0}$ of f_0 becomes the search direction $\bar{\mathbf{y}}_g$ of Eq. (3.2.1). Moreover, even when obtaining the search direction $\bar{\mathbf{y}}_g$ in the case when any of the inequality constraint conditions become active, as will be shown in Sect. 3.7 and beyond, the search direction $\bar{\mathbf{y}}_g$ satisfying the inequality constraints can be obtained using the descent direction $\bar{\mathbf{y}}_{g0}$ of the objective function f_0 and the descent direction $\bar{\mathbf{y}}_{gi}$ of each the active constraint functions f_i . The gradient method is used in this case too.

In this book, from Sects. 3.3 to 3.5, unconstrained problems will be considered. Here, for simplicity, f_i , \mathbf{g}_i and $\bar{\mathbf{y}}_{gi}$ are written as f , \mathbf{g} and $\bar{\mathbf{y}}_g$, respectively.

Before we proceed further, let us define the symbols while referring to Figs. 3.1 and 3.2. For each $k \in \mathbb{N}$, let $\mathbf{x}_k \in X$ be a trial point and \mathbf{g} be the gradient of f at \mathbf{x}_k not identical to the zero vector $\mathbf{0}_X$. We suppose that \mathbf{g} is known in advance and then seek the direction $\bar{\mathbf{y}}_g \in X$ in which f decreases.

Fig. 3.1 Definition of gradient \mathbf{g}

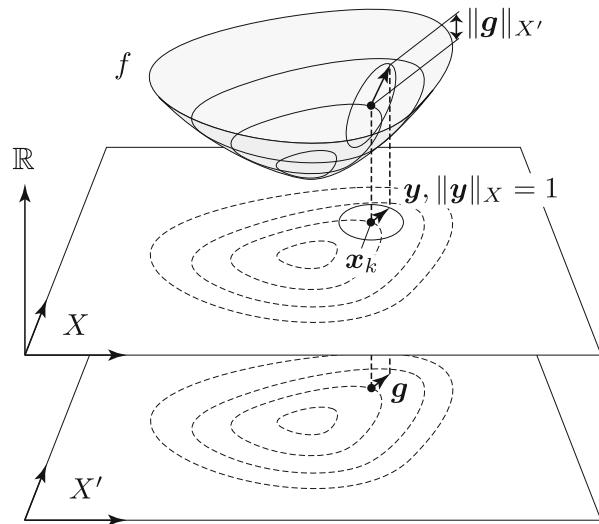
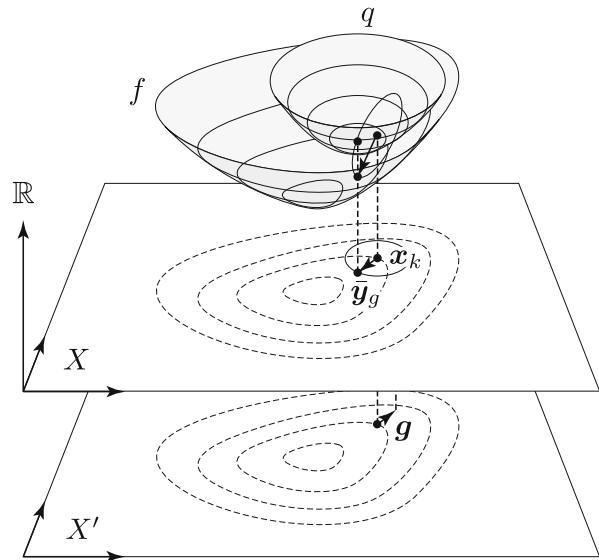


Fig. 3.2 Gradient method



At this point let us confirm the meaning of \mathbf{g} . Consider the Taylor expansion of f around \mathbf{x}_k given by

$$f(\mathbf{x}_k + \mathbf{y}) = f(\mathbf{x}_k) + \mathbf{g} \cdot \mathbf{y} + o(\|\mathbf{y}\|_X). \quad (3.3.1)$$

Here, if the definition of a Fréchet derivative (see Definition 4.5.4 in Chap. 4) is used, then \mathbf{g} is an element of the dual space X' (Definition 4.4.5) of X and the size (norm) of \mathbf{g} is defined by

$$\|\mathbf{g}\|_{X'} = \max_{\mathbf{y} \in X} \frac{|\langle \mathbf{g}, \mathbf{y} \rangle_{X' \times X}|}{\|\mathbf{y}\|_X} = \max_{\mathbf{y} \in X, \|\mathbf{y}\|_X=1} |\langle \mathbf{g}, \mathbf{y} \rangle_{X' \times X}|.$$

If $X = \mathbb{R}^d$, then $X' = \mathbb{R}^d$ and the dual product is given by $\langle \mathbf{g}, \mathbf{y} \rangle_{X' \times X} = \mathbf{g} \cdot \mathbf{y}$. Based on this definition, $\|\mathbf{g}\|_{X'}$ represents the maximum value of $|\mathbf{g} \cdot \mathbf{y}|$ over all the direction $\mathbf{y} \in X$ such that $\|\mathbf{y}\|_X = 1$. Moreover, the direction of \mathbf{g} is perpendicular with respect to the contour lines of f . This is because in Eq. (3.3.1), with $\mathbf{x}_k + \mathbf{y} \in X$ as a point on the contour line and \mathbf{y} is taken to be a sufficiently small vector,

$$\mathbf{g} \cdot \mathbf{y} \approx f(\mathbf{x}_k + \mathbf{y}) - f(\mathbf{x}_k) = 0$$

holds. Figure 3.1 illustrates this relationship.

From these relationships, we infer that \mathbf{g} points in the direction such that f increases the most. Hence, from the fact that $X = X'$ holds in a finite-dimensional vector space (Sect. 4.4.6), if $\bar{\mathbf{y}}_g \in X$ is chosen such that

$$\bar{\mathbf{y}}_g = -\mathbf{g}, \quad (3.3.2)$$

then we get

$$f(\mathbf{x}_k + \bar{\mathbf{y}}_g) - f(\mathbf{x}_k) = -\|\bar{\mathbf{y}}_g\|_X^2 + o(\|\bar{\mathbf{y}}_g\|_X).$$

Here, if $\|\bar{\mathbf{y}}_g\|_X$ is sufficiently small, f decreases.

Let us generalize this method. The method for obtaining the descent direction $\bar{\mathbf{y}}_g \in X$ as the solution to the following problem is called the gradient method.

Problem 3.3.1 (Gradient Method) Let $X = \mathbb{R}^d$ and let $A \in \mathbb{R}^{d \times d}$ be a positive definite real symmetric matrix (Definition 2.4.5). Let the gradient of f at $\mathbf{x}_k \in X$ which is not a local minimum point with respect to $f \in C^1(X; \mathbb{R})$ be $\mathbf{g}(\mathbf{x}_k) \in X' = \mathbb{R}^d$. In this case, obtain $\bar{\mathbf{y}}_g \in X$ which satisfies

$$\bar{\mathbf{y}}_g \cdot (A\mathbf{y}) = -\mathbf{g}(\mathbf{x}_k) \cdot \mathbf{y} \quad (3.3.3)$$

with respect to an arbitrary $\mathbf{y} \in X$. □

Equation (3.3.3) is an expression using the inner product with an arbitrary $y \in X$. This equation is equivalent to obtaining \bar{y}_g via

$$\bar{y}_g = -A^{-1}g. \quad (3.3.4)$$

The reason for using the inner product is so that when defining the gradient method in function space in Chap. 7, it becomes a natural extension of Problem 3.3.1. Moreover, Eq. (3.3.2) is the gradient method in the case when A is the identity matrix I . Let us confirm the fact that the solution \bar{y}_g of Problem 3.3.1 reduces f with the following theorem.

Theorem 3.3.2 (Gradient Method) *The solution \bar{y}_g of Problem 3.3.1 is the descent direction of f at x_k .* \square

Proof Since A is a positive definite symmetric matrix, a positive constant α exists and the inequality

$$y \cdot (Ay) \geq \alpha \|y\|_X^2, \quad A = A^\top,$$

holds with respect to an arbitrary $y \in X$. This relationship and Eq. (3.3.3) can be used to establish

$$\begin{aligned} f(x_k + \bar{\epsilon} \bar{y}_g) - f(x_k) &= \bar{\epsilon} g \cdot \bar{y}_g + o(\bar{\epsilon}) = -\bar{\epsilon} \bar{y}_g \cdot (A \bar{y}_g) + o(\bar{\epsilon}) \\ &\leq -\bar{\epsilon} \alpha \|\bar{y}_g\|_X^2 + o(\bar{\epsilon}) \end{aligned}$$

with respect to the positive constant $\bar{\epsilon}$. Consequently, if $\bar{\epsilon}$ is sufficiently small, then f decreases in value. \square

Let us define the descent direction's descent angle as follows.

Definition 3.3.3 (Descent Angle) For $x_k \in X$, $g \in X'$ is taken to be the gradient and $\bar{y}_g \in X$ the descent direction. In this case, $\theta \in [0, \pi]$ is defined by

$$\cos \theta = -\frac{\langle g, \bar{y}_g \rangle_{X' \times X}}{\|g\|_{X'} \|\bar{y}_g\|_X}$$

is called the descent angle of \bar{y}_g at x_k . \square

If A is set to be the identity matrix I in the gradient method (Problem 3.3.1), the descent angle θ of \bar{y}_g vanishes. This iterative method is called the maximum descent method. However, it is not necessarily the case that convergence is faster with the maximum descent method. This will become clear through comparison with the conjugate gradient method (Problem 3.4.9) which will be discussed later in this section.

Let us consider the geometric meaning of $\bar{\mathbf{y}}_g$ obtained by the gradient method. Problem 3.3.1 is equivalent to seeking $\bar{\mathbf{y}}_g \in X$ which satisfies

$$q(\bar{\mathbf{y}}_g) = \min_{\mathbf{y} \in X} \left\{ q(\mathbf{y}) = \frac{1}{2} \mathbf{y} \cdot (\mathbf{A} \mathbf{y}) + \mathbf{g} \cdot \mathbf{y} + f(\mathbf{x}_k) \right\}. \quad (3.3.5)$$

In fact, if the condition $q'(\bar{\mathbf{y}}_g)[\mathbf{y}] = 0$ holds true for any $\mathbf{y} \in X$, then so does Eq. (3.3.3) and vice versa. Figure 3.2 shows the function q in this case. Here, q is an elliptic paraboloid and its minimum point is $\mathbf{x}_k + \bar{\mathbf{y}}_g$. The size of $\bar{\mathbf{y}}_g$ depends on the choice of \mathbf{A} . Therefore, if one wants the step size $\|\mathbf{y}_g\|_X = \|\bar{\epsilon}_g \bar{\mathbf{y}}_g\|_X$ to be ϵ_g , the following calculation should be carried out. Introduce a positive constant c_a as an adjustment parameter and change Eq. (3.3.4) to

$$\mathbf{y}_g = -(c_a \mathbf{A})^{-1} \mathbf{g}. \quad (3.3.6)$$

It should be noted here that if c_a is made bigger, the size of \mathbf{y}_g becomes smaller. Hence, when the step size ϵ_g and the solution $\bar{\mathbf{y}}_g = \bar{\mathbf{y}}_g(\mathbf{x}_0)$ of the gradient method (Problem 3.3.1) at initial point \mathbf{x}_0 are given, we have

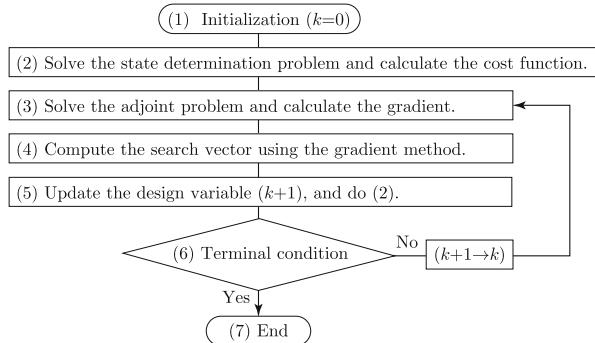
$$c_a = \frac{\|\bar{\mathbf{y}}_g\|_X}{\epsilon_g}. \quad (3.3.7)$$

Here we assume that c_a is obtained in this way at the initial step ($k = 0$) and its value is used in order to obtain the search vector via Eq. (3.3.6) at the succeeding steps ($k \in \mathbb{N}$) too. Then the step size $\|\mathbf{y}_{gk}\|_X$ roughly takes the size ϵ_g for a while. Moreover, the step size becomes zero when the trial point approaches the point of convergence. In this case, it is equivalent to seeking $\mathbf{y}_g \in X$ which satisfies

$$q(\mathbf{y}_g) = \min_{\mathbf{y} \in X} \left\{ q(\mathbf{y}) = \frac{1}{2} \mathbf{y} \cdot (c_a \mathbf{A} \mathbf{y}) + \mathbf{g} \cdot \mathbf{y} + f(\mathbf{x}_k) \right\}. \quad (3.3.8)$$

In the above equation, we emphasize that the magnitude of c_a depends on the choice of the free parameter ϵ_g . So, obtaining an appropriate value for c_a is clearly not straightforward. Nevertheless, if we can determine the step size, then we can use Eq. (3.3.7) to decide for c_a . In the case of domain variations, discussed in Chap. 9, for instance, the magnitude of a domain variation (step size) is defined by a norm of the strain with respect to the domain variation, such as the square root of the integral of squared strain or the maximum strain. In such a situation, we can imagine that a value of 0.05 for ϵ_g would be a good choice. However, if we do not have any idea about the step size, then we have to consider another way to decide the value of c_a .

Fig. 3.3 Algorithm of gradient method



One possible way to determine a good choice for c_a is to assume that the objective function reduces at some rate after a domain variation.¹ We illustrate this method as follows. Suppose that the objective function $f(\mathbf{x}_0)$ and the gradient $\mathbf{g}(\mathbf{x}_0)$ at $k = 0$ are given, and a search vector $\bar{\mathbf{y}}_g$ is obtained by the gradient method. Also, let us assume that the objective function reduces at a rate of $\alpha \in (0, 1)$ after every domain variation. Then, we have the estimate

$$f(\mathbf{x}_0 + \bar{\epsilon}_g \bar{\mathbf{y}}_g) - f(\mathbf{x}_0) \approx \alpha f(\mathbf{x}_0) \approx \bar{\epsilon}_g \mathbf{g}(\mathbf{x}_0) \cdot \bar{\mathbf{y}}_g.$$

When ‘ \approx ’ is considered ‘ $=$ ’, c_a is given by

$$c_a = \frac{1}{\bar{\epsilon}_g} = \frac{\mathbf{g}(\mathbf{x}_0) \cdot \bar{\mathbf{y}}_g}{\alpha f(\mathbf{x}_0)}. \quad (3.3.9)$$

Based on the observations above, let us develop a simple algorithm based on the gradient method. In this chapter, we shall make use of some particular statements when stating the steps of the algorithms. More precisely, with the supposition that optimal design problems may be solved, the following expressions will be used. The phrase ‘Calculate $f(\mathbf{x}_k)$ ’ will be written as ‘Solve the state determination problem and calculate $f(\mathbf{x}_k)$ ’. Furthermore, we will write ‘Calculate $\mathbf{g}(\mathbf{x}_k)$ ’ as ‘Solve the adjoint problem with respect to f and calculate $\mathbf{g}(\mathbf{x}_k)$ ’. The reason for these is because the calculation becomes like that in an optimal design problem, as explained at the start of Sect. 3.1.

With this background in mind, we now provide examples of algorithms using the gradient method, starting with the simplest one. The adjustment parameter for determining the step size c_a is assumed to be given in advance. Figure 3.3 illustrates an overview of the method.

Algorithm 3.1 (Gradient Method) In Problem 3.1.1, f_0 is denoted by f and all inequality constraints are assumed to be inactive.

¹Julius Fergy T. Rabago (private communication).

- (1) Define the following parameters: initial point \mathbf{x}_0 , positive definite symmetric matrix \mathbf{A} (\mathbf{I} if there is no particular specification), positive constant for adjusting the step size c_a and positive constant ϵ_0 needed for the convergence check. Set $k = 0$.
- (2) Solve the state determination problem and calculate $f(\mathbf{x}_k)$.
- (3) Solve the adjoint problem with respect to f and calculate $\mathbf{g}(\mathbf{x}_k)$.
- (4) Calculate \mathbf{y}_g by Eq. (3.3.6).
- (5) Let $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{y}_g$. Solve the state determination problem and calculate $f(\mathbf{x}_{k+1})$.
- (6) Check the final condition $|f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)| \leq \epsilon_0$.
 - Proceed to (7) if the final condition is satisfied.
 - Otherwise, substitute $k + 1$ into k and return to (3).
- (7) Complete the calculation. □

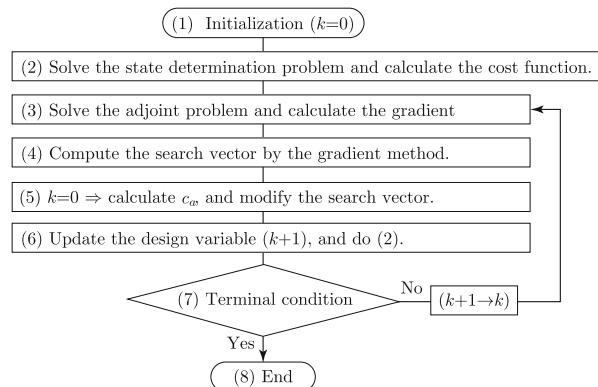
In Algorithm 3.1, $|f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)| \leq \epsilon_0$ was used as a stopping criterion. Other than this, conditions such as k is not over an upper limit or $\|\mathbf{y}_g\|_X \leq \epsilon_0$ in which attention is given to the variation of the design variable can also be utilized.

If one wanted c_a to be determined so that the first step size is a specified ϵ_g , the following algorithm can be used. Figure 3.4 shows an overview of its steps.

Algorithm 3.2 (Gradient Method with Initial Step Size Specified) In Problem 3.1.1, f_0 is denoted by f and all inequality constraints are assumed to be inactive.

- (1) Define the following parameters: initial point \mathbf{x}_0 , positive definite symmetric matrix \mathbf{A} , positive constant for the initial step-size ϵ_g and positive constant ϵ_0 needed for the convergence check. Let $c_a = 1$ and set $k = 0$.
- (2) Solve the state determination problem and calculate $f(\mathbf{x}_k)$.
- (3) Solve the adjoint problem with respect to f and calculate $\mathbf{g}(\mathbf{x}_k)$.
- (4) Use Eq. (3.3.6) to calculate \mathbf{y}_g .

Fig. 3.4 Algorithm of gradient method when the initial value of the step size is given



- (5) When $k = 0$, let $\mathbf{y}_g = \bar{\mathbf{y}}_g$ and obtain c_a using Eq. (3.3.7). Moreover, substitute $\bar{\mathbf{y}}_g/c_a$ into \mathbf{y}_g .
- (6) Let $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{y}_g$. Solve the state determination problem and calculate $f(\mathbf{x}_{k+1})$.
- (7) Check the final condition $|f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)| \leq \epsilon_0$.
 - Proceed to (8) when the terminal condition is satisfied.
 - Otherwise, substitute $k + 1$ into k and return to (3).
- (8) Complete the calculation. □

3.4 Step Size Criterion

Next let us consider a method for appropriately deciding the step size $\|\mathbf{y}_g\|_X$.

If the search direction $\bar{\mathbf{y}}_g$ is already known, the variable in an optimization problem is only $\bar{\epsilon}_g$ of Eq. (3.2.1). Hence, an optimization problem such as the following with $\bar{\epsilon}_g$ taken to be a design variable can be considered to suggest a method for determining the step size $\|\bar{\epsilon}_g \bar{\mathbf{y}}_g\|_X$ from its solutions. This method is called the strict line search method.

Problem 3.4.1 (Strict Line Search Method) Let $X = \mathbb{R}^d$. Given $f \in C^1(X; \mathbb{R})$, $\mathbf{x}_k \in X$ and $\bar{\mathbf{y}}_g \in X$, obtain $\bar{\epsilon}_g$ which satisfies

$$\min_{\bar{\epsilon}_g \in (0, \infty)} f(\mathbf{x}_k + \bar{\epsilon}_g \bar{\mathbf{y}}_g). \quad \square$$

The algorithm for solving Problem 3.4.1 employs methods to solve non-linear equations. For example, a method such as

- Bisection method
- Secant method

can be considered. The bisection method repeats operations for finding the midpoint of the region such as $\bar{\epsilon}_g$ in which f changes from decreasing to increasing. In this case, the convergence order to an exact solution for $\bar{\epsilon}_g$ is one. When using the secant method, it is viewed as a problem obtaining $\bar{\epsilon}_g$ such that the gradient of f with respect to $\bar{\epsilon}_g$ is zero. To solve the problem, the updating equation of the Newton–Raphson method, which will be shown later, is used. However, in the secant method, the gradient of f is replaced with the difference (Practices 3.1 and 3.2). It is known that the convergence order of this method is the golden ratio $(1 + \sqrt{5})/2$.

The following results can be obtained regarding the convergence of an exact solution of \mathbf{x}_k computed through the strict line search method. Assume that the cost function is a second-order function. Suppose that the search direction is obtained via the maximum descent method. In this case, the step size obtained from the strict line search method is the solution of the following problem.

Problem 3.4.2 (Strict Line Search of 2nd-Order Optimization Problem) Let $X = \mathbb{R}^d$. Suppose $\mathbf{B} \in \mathbb{R}^{d \times d}$ is a positive definite symmetric matrix, $\mathbf{b} \in X$ is a given vector, and

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x} \cdot (\mathbf{B}\mathbf{x}) + \mathbf{b} \cdot \mathbf{x} \quad (3.4.1)$$

is a cost function. Given these assumptions, find $\bar{\mathbf{y}}_g \in X$ using the maximum descent method with respect to $\mathbf{x}_k \in X$ and obtain $\bar{\epsilon}_g \in (0, \infty)$ which satisfies Problem 3.4.1. \square

Answer If $f(\mathbf{x}_k + \bar{\epsilon}_g \bar{\mathbf{y}}_g)$ is written as $\bar{f}(\bar{\epsilon}_g)$, we can write

$$\begin{aligned} \bar{f}(\bar{\epsilon}_g) &= \frac{1}{2}(\mathbf{x}_k + \bar{\epsilon}_g \bar{\mathbf{y}}_g) \cdot \{\mathbf{B}(\mathbf{x}_k + \bar{\epsilon}_g \bar{\mathbf{y}}_g)\} + \mathbf{b} \cdot (\mathbf{x}_k + \bar{\epsilon}_g \bar{\mathbf{y}}_g) \\ &= \bar{\epsilon}_g^2 \frac{1}{2} \bar{\mathbf{y}}_g \cdot (\mathbf{B} \bar{\mathbf{y}}_g) + \bar{\epsilon}_g \bar{\mathbf{y}}_g \cdot \mathbf{g} + f(\mathbf{x}_k), \end{aligned}$$

where $\mathbf{g} = \mathbf{B}\mathbf{x}_k + \mathbf{b}$ was used in the second equality. In view of the strict line search method, the equation

$$\frac{d\bar{f}}{d\bar{\epsilon}_g} = \bar{\epsilon}_g \bar{\mathbf{y}}_g \cdot (\mathbf{B} \bar{\mathbf{y}}_g) + \bar{\mathbf{y}}_g \cdot \mathbf{g} = 0$$

yields

$$\bar{\epsilon}_g = -\frac{\bar{\mathbf{y}}_g \cdot \mathbf{g}}{\bar{\mathbf{y}}_g \cdot (\mathbf{B} \bar{\mathbf{y}}_g)}.$$

Furthermore, if $\bar{\mathbf{y}}_g$ is the solution of the maximum descent method, then one has $\bar{\mathbf{y}}_g = -\mathbf{g}$, which gives

$$\bar{\epsilon}_g = -\frac{\bar{\mathbf{y}}_g \cdot \mathbf{g}}{\bar{\mathbf{y}}_g \cdot (\mathbf{B} \bar{\mathbf{y}}_g)} = \frac{\mathbf{g} \cdot \mathbf{g}}{\mathbf{g} \cdot (\mathbf{B} \mathbf{g})} = \frac{\mathbf{g} \cdot \mathbf{g}}{\bar{\mathbf{y}}_g \cdot (\mathbf{B} \bar{\mathbf{y}}_g)}. \quad (3.4.2)$$

\square

In this way, the strict line search method can be used to provide the following results regarding the convergence when the iterative method is repeated while seeking $\bar{\epsilon}_g$.

Theorem 3.4.3 (Convergence of the Strict Line Search Method) *The sequence of iterates $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$ formed by the iterative method using the solution to Problem 3.4.2, $\bar{\mathbf{y}}_g \in X$, and $\bar{\epsilon}_g$ satisfies*

$$\|\mathbf{x}_{k+1} - \mathbf{x}\|_{\mathbf{B}} \leq \left| \frac{\lambda_d - \lambda_1}{\lambda_1 + \lambda_d} \right| \|\mathbf{x}_k - \mathbf{x}\|_{\mathbf{B}},$$

where \mathbf{x} is a local minimum point, λ_1 and λ_d denote the minimum and maximum eigenvalues, respectively, and $\|\mathbf{x}\|_{\mathbf{B}} = \sqrt{\mathbf{x} \cdot (\mathbf{B}\mathbf{x})}$. \square

Proof The objective function in Problem 3.4.2 can be written as

$$f(\mathbf{x}) = \frac{1}{2} (\mathbf{x} + \mathbf{B}^{-1}\mathbf{b}) \cdot \left\{ \mathbf{B} (\mathbf{x} + \mathbf{B}^{-1}\mathbf{b}) \right\} - \frac{1}{2} \mathbf{b} \cdot (\mathbf{B}^{-1}\mathbf{b}).$$

Observe that even if $\mathbf{x} + \mathbf{B}^{-1}\mathbf{b}$ is replaced by \mathbf{x} , the evaluation of $\mathbf{x}_{k+1} - \mathbf{x}$ remains unchanged. Moreover, since the second term on the right-hand side in the above equation is independent of \mathbf{x} , then even if it is omitted, the evaluation of $\mathbf{x}_{k+1} - \mathbf{x}$ does not change. Therefore, it suffices to consider the problem of finding the minimum point of

$$\bar{f}(\mathbf{x}) = \frac{1}{2} \mathbf{x} \cdot (\mathbf{B}\mathbf{x}).$$

When $\mathbf{g}_k = \mathbf{g}(\mathbf{x}_k) = \mathbf{B}\mathbf{x}_k$, the maximum descent method can be used to obtain $\bar{\mathbf{y}}_k = -\mathbf{g}_k$. Furthermore, as a result of the strict line search method, Eq. (3.4.2) can be used to obtain

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \frac{\mathbf{g}_k \cdot \mathbf{g}_k}{\mathbf{g}_k \cdot (\mathbf{B}\mathbf{g}_k)} \bar{\mathbf{y}}_k$$

to form the point sequence. In this case,

$$\begin{aligned} \bar{f}(\mathbf{x}_{k+1}) &= \frac{1}{2} \left(\mathbf{x}_k - \frac{\mathbf{g}_k \cdot \mathbf{g}_k}{\mathbf{g}_k \cdot (\mathbf{B}\mathbf{g}_k)} \mathbf{g}_k \right) \cdot \left\{ \mathbf{B} \left(\mathbf{x}_k - \frac{\mathbf{g}_k \cdot \mathbf{g}_k}{\mathbf{g}_k \cdot \mathbf{B}\mathbf{g}_k} \mathbf{g}_k \right) \right\} \\ &= \frac{1}{2} \left[\mathbf{x}_k \cdot (\mathbf{B}\mathbf{x}_k) - \frac{2(\mathbf{g}_k \cdot \mathbf{g}_k) \{ \mathbf{g}_k \cdot (\mathbf{B}\mathbf{x}_k) \} - (\mathbf{g}_k \cdot \mathbf{g}_k)^2}{\mathbf{g}_k \cdot (\mathbf{B}\mathbf{g}_k)} \right] \\ &= \frac{1}{2} \left\{ \mathbf{x}_k \cdot (\mathbf{B}\mathbf{x}_k) - \frac{(\mathbf{g}_k \cdot \mathbf{g}_k)^2}{\mathbf{g}_k \cdot (\mathbf{B}\mathbf{g}_k)} \right\} \\ &= \frac{1}{2} \mathbf{x}_k \cdot (\mathbf{B}\mathbf{x}_k) \left(1 - \frac{(\mathbf{g}_k \cdot \mathbf{g}_k)^2}{\{ \mathbf{x}_k \cdot (\mathbf{B}\mathbf{x}_k) \} \{ \mathbf{g}_k \cdot (\mathbf{B}\mathbf{g}_k) \}} \right) \\ &= \left(1 - \frac{(\mathbf{g}_k \cdot \mathbf{g}_k)^2}{\{ \mathbf{g}_k \cdot (\mathbf{B}^{-1}\mathbf{g}_k) \} \{ \mathbf{g}_k \cdot (\mathbf{B}\mathbf{g}_k) \}} \right) \bar{f}(\mathbf{x}_k) \end{aligned}$$

is established. Since \mathbf{B} is positive definite, we can apply Kantorovich's inequality to obtain

$$\frac{4\lambda_1\lambda_d}{(\lambda_1 + \lambda_d)^2} \leq \frac{(\mathbf{y} \cdot \mathbf{y})^2}{\{\mathbf{y} \cdot (\mathbf{B}^{-1}\mathbf{y})\} \{\mathbf{y} \cdot (\mathbf{B}\mathbf{y})\}},$$

for any $\mathbf{y} \in X$. Therefore,

$$\bar{f}(\mathbf{x}_{k+1}) \leq \left(1 - \frac{4\lambda_1\lambda_d}{(\lambda_1 + \lambda_d)^2}\right) \bar{f}(\mathbf{x}_k) = \left(\frac{\lambda_d - \lambda_1}{\lambda_1 + \lambda_d}\right)^2 \bar{f}(\mathbf{x}_k),$$

and thus, the desired result. \square

Let us consider the characteristics of the strict line search method with the above results in mind. The strict line search method requires the minimization problem with only the step size as a design variable to be solved accurately. To do this, iterative algorithms such as the bisection method or the secant method are of great practical importance, and thus become necessary. For problems where the calculation of the gradient is rather more difficult compared to the calculation of the cost function with respect to the design variable, it is considered that an effective algorithm can be formulated such that the calculation of the gradient is unnecessary, like the bisection method. However, once the design variable moves in the search direction, the gradient changes and the search direction determined by the gradient method also changes. Even in such situations, it is considered that it is not necessarily a good idea to seek the minimum point accurately using the old search direction. In particular, in the case when using an algorithm in which there is a need for the recalculation of the gradient after $\bar{\epsilon}_g$ is updated, as with the secant method (Practice 3.2), it is considered that updating the search direction via the gradient method would improve convergence rather than just continuing to use the old search direction.

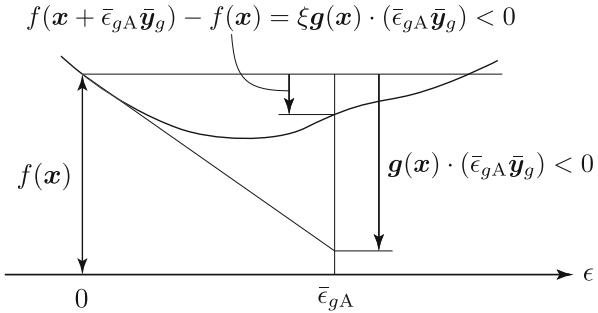
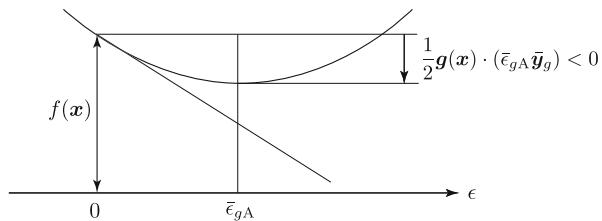
In the sequel, we shall examine a method focusing on the range over which the non-linearity and gradient of the cost function is effective without worrying about whether it is strictly the case. The criterion shown below provides the upper and lower limit of the step size. With respect to the upper limit of the step size $\|\bar{\epsilon}_g \bar{\mathbf{y}}_g\|_X$, conditions such as the one that follows are known [5].

Definition 3.4.4 (Armijo Criterion) Suppose $\mathbf{g}(\mathbf{x}_k)$ is the gradient of $f(\mathbf{x}_k)$, $\bar{\mathbf{y}}_g$ is the search direction and $\xi \in (0, 1)$ is the parameter adjusting the upper limit of the step size. If

$$f(\mathbf{x}_k + \bar{\epsilon}_g \bar{\mathbf{y}}_g) - f(\mathbf{x}_k) \leq \xi \mathbf{g}(\mathbf{x}_k) \cdot (\bar{\epsilon}_g \bar{\mathbf{y}}_g) < 0 \quad (3.4.3)$$

holds for any $\bar{\epsilon}_g > 0$, $\bar{\epsilon}_g$ satisfies the Armijo criterion. \square

If the upper limit of $\bar{\epsilon}_g$ satisfying the Armijo criterion is written as $\bar{\epsilon}_{gA}$, a relationship such as shown in Fig. 3.5 is established. The left-hand side of Eq. (3.4.3)

Fig. 3.5 Armijo criterion**Fig. 3.6** Armijo criterion with respect to a second-order function

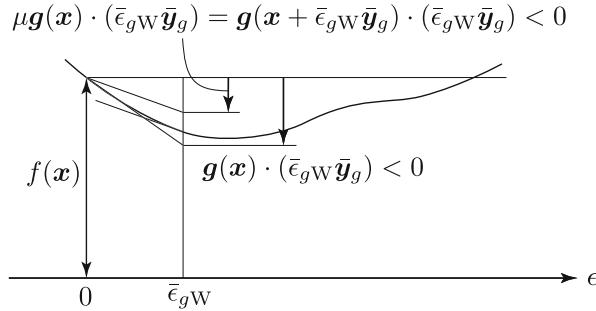
takes a negative value by which the non-linear function f actually reduces when \mathbf{x} fluctuates from \mathbf{x}_k by just $\bar{\epsilon}_g \bar{\mathbf{y}}_g$. In contrast, the quantity $\mathbf{g}(\mathbf{x}_k) \cdot (\bar{\epsilon}_g \bar{\mathbf{y}}_g)$ on the right-hand side admits a negative value when the reduction in f is estimated using the gradient. Equality holds when $\bar{\epsilon}_g$ is sufficiently small. However, the case is different when $\bar{\epsilon}_g$ is of a certain size. Instances when $\xi \in (0, 1)$ provides the ratio by which this difference is permitted. Making ξ close to unity represents the fact that their difference is not allowed and making ξ close to zero represents that their difference is permitted. Therefore, the Armijo criterion in effect provides the condition for deciding the step size at a level such that the estimated value using the gradient of f is not too far away from the actual value of reduction. It must be noted that ξ is restricted on the unit interval $(0, 1)$ since the Armijo criterion actually fails when $\xi > 1$.

Moreover, the following results can be used as a benchmark for ξ . If $f(\mathbf{x})$ is a second-order function and the upper limit of the Armijo criterion when $\xi = 1/2$ is $\bar{\epsilon}_{gA}$, $\mathbf{x}_k + \bar{\epsilon}_{gA} \bar{\mathbf{y}}_g$ becomes the minimum point of f , see Fig. 3.6. In fact,

$$f(\mathbf{x}_k + \bar{\epsilon}_{gA} \bar{\mathbf{y}}_g) - f(\mathbf{x}_k) = \mathbf{g}(\mathbf{x}_k) \cdot (\bar{\epsilon}_{gA} \bar{\mathbf{y}}_g) + \frac{1}{2} (\bar{\epsilon}_{gA} \bar{\mathbf{y}}_g) \cdot (\mathbf{B}(\bar{\epsilon}_{gA} \bar{\mathbf{y}}_g)) \quad (3.4.4)$$

is established with respect to the second-order function of Eq. (3.4.1). Here, if $\mathbf{x}_k + \bar{\epsilon}_{gA} \bar{\mathbf{y}}_g$ is a local minimum point, the Taylor expansion of $\mathbf{g}(\mathbf{x}_k)$ is given by

$$\mathbf{g}(\mathbf{x}_k + \bar{\epsilon}_{gA} \bar{\mathbf{y}}_g) = \mathbf{g}(\mathbf{x}_k) + \mathbf{B}(\bar{\epsilon}_{gA} \bar{\mathbf{y}}_g) = \mathbf{0}_{X'}. \quad (3.4.5)$$

Fig. 3.7 Wolfe criterion

If Eq. (3.4.5) is substituted into Eq. (3.4.4), then one obtains

$$f(\mathbf{x}_k + \bar{\epsilon}_{gA} \bar{\mathbf{y}}_g) - f(\mathbf{x}_k) = \frac{1}{2} \mathbf{g}(\mathbf{x}_k) \cdot (\bar{\epsilon}_{gA} \bar{\mathbf{y}}_g).$$

On the other hand, conditions such as the following are known for providing the lower limit for the step size $\|\bar{\epsilon}_g \bar{\mathbf{y}}_g\|_X$ [172].²

Definition 3.4.5 (Wolfe Criterion) Let $\mathbf{g}(\mathbf{x}_k)$ be the gradient of $f(\mathbf{x}_k)$, $\bar{\mathbf{y}}_g$ the search direction, $\xi \in (0, 1)$ the parameter used in the Armijo criterion, $\mu \in (0, 1)$ the parameter which adjusts the lower limit of the step size and suppose that $0 < \xi < \mu < 1$ is satisfied. In this case, if

$$\mu \mathbf{g}(\mathbf{x}_k) \cdot (\bar{\epsilon}_g \bar{\mathbf{y}}_g) \leq \mathbf{g}(\mathbf{x}_k + \bar{\epsilon}_g \bar{\mathbf{y}}_g) \cdot (\bar{\epsilon}_g \bar{\mathbf{y}}_g) < 0 \quad (3.4.6)$$

holds with respect to $\bar{\epsilon}_g > 0$, $\bar{\epsilon}_g$ is said to satisfy the Wolfe criterion. \square

If the lower limit value of $\bar{\epsilon}_g$ which satisfies the Wolfe criterion is written as $\bar{\epsilon}_{gW}$, a geometric relationship such as the one shown in Fig. 3.7 is established. The term $\mathbf{g}(\mathbf{x}_k)$ on the left-hand side of Eq. (3.4.6) represents the gradient of f at \mathbf{x}_k . On the other hand, the expression $\mathbf{g}(\mathbf{x}_k + \bar{\epsilon}_g \bar{\mathbf{y}}_g)$ on the right-hand side represents the gradient of f obtained when moving \mathbf{x} from \mathbf{x}_k to $\mathbf{x}_k + \bar{\epsilon}_g \bar{\mathbf{y}}_g$. The following observations regarding the Wolfe criterion are established:

- (1) If the condition $\mathbf{g}(\mathbf{x}_k + \bar{\epsilon}_g \bar{\mathbf{y}}_g) \cdot (\bar{\epsilon}_g \bar{\mathbf{y}}_g) \leq \mathbf{g}(\mathbf{x}_k) \cdot (\bar{\epsilon}_g \bar{\mathbf{y}}_g) < 0$ holds for some $\bar{\epsilon}_g > 0$, then there is no $\bar{\epsilon}_g > 0$ such that Eq. (3.4.6) holds. This condition expresses the fact that when \mathbf{x} moves in the direction of $\bar{\mathbf{y}}_g$, the negative gradient which would reduce f becomes an even greater negative gradient. It shows that in such a case there is no need to provide a lower limit to the step size.
- (2) If, for some $\bar{\epsilon}_g > 0$, $\mathbf{g}(\mathbf{x}_k) \cdot (\bar{\epsilon}_g \bar{\mathbf{y}}_g) < \mathbf{g}(\mathbf{x}_k + \bar{\epsilon}_g \bar{\mathbf{y}}_g) \cdot (\bar{\epsilon}_g \bar{\mathbf{y}}_g) < 0$ holds, $\bar{\epsilon}_g > 0$ exists such that Eq. (3.4.6) holds. This condition shows, in contrast to

²In the literature, the Armijo criterion is in fact a special case of the Wolfe criterion, but in this book, only the condition which gives the lower limit of $\bar{\epsilon}_g$ is referred to as the Wolfe criterion.

(1) above, that the gradient decreases when \mathbf{x} moves in the direction of $\bar{\mathbf{y}}_g$. If μ is made smaller than one in Eq. (3.4.6), the lower limit $\bar{\epsilon}_g w$ of $\bar{\epsilon}_g$ becomes bigger.

Hence, the Wolfe criterion is a condition which ideally requires the step size to be large enough such that the validity of the gradient is lost to around the proportion of μ .

Meanwhile, the requirement $\xi < \mu$ is based on the following observations. The Taylor expansion of f about $\mathbf{x}_k + \bar{\epsilon}_g \bar{\mathbf{y}}_g$ is written as

$$f(\mathbf{x}_k) = f(\mathbf{x}_k + \bar{\epsilon}_g \bar{\mathbf{y}}_g) - \mathbf{g}(\mathbf{x}_k + \bar{\epsilon}_g \bar{\mathbf{y}}_g) \cdot (\bar{\epsilon}_g \bar{\mathbf{y}}_g) + o(\bar{\epsilon}_g).$$

In this case, if the Wolfe criterion is satisfied, then the following relations hold:

$$\begin{aligned} \mu \mathbf{g}(\mathbf{x}_k) \cdot (\bar{\epsilon}_g \bar{\mathbf{y}}_g) - o(\bar{\epsilon}_g) &\leq \mathbf{g}(\mathbf{x}_k + \bar{\epsilon}_g \bar{\mathbf{y}}_g) \cdot (\bar{\epsilon}_g \bar{\mathbf{y}}_g) - o(\bar{\epsilon}_g) \\ &= f(\mathbf{x}_k + \bar{\epsilon}_g \bar{\mathbf{y}}_g) - f(\mathbf{x}_k). \end{aligned}$$

On the other hand, if the Armijo criterion is satisfied, then we have that

$$f(\mathbf{x}_k + \bar{\epsilon}_g \bar{\mathbf{y}}_g) - f(\mathbf{x}_k) \leq \xi \mathbf{g}(\mathbf{x}_k) \cdot (\bar{\epsilon}_g \bar{\mathbf{y}}_g).$$

Hence, if both conditions are satisfied, then the following requirement must hold:

$$(\mu - \xi) \mathbf{g}(\mathbf{x}_k) \cdot (\bar{\epsilon}_g \bar{\mathbf{y}}_g) \leq o(\bar{\epsilon}_g).$$

Here, the fact that $\mathbf{g}(\mathbf{x}_k) \cdot (\bar{\epsilon}_g \bar{\mathbf{y}}_g) \leq 0$ implies that if the right-hand side is positive, or else the absolute value is sufficiently small, then the inequality holds when $\xi < \mu$.

An example of an algorithm in which the step size is controlled so that the Armijo criterion and the Wolfe criterion are satisfied is shown below. Figure 3.8 provides an overview of the steps in the algorithm.

Algorithm 3.3 (Armijo Criterion and Wolfe Criterion) Consider Problem 3.1.1. Let f_0 be f and all inequality constraints be inactive.

- (1) Define the following parameters: the initial point \mathbf{x}_0 , positive definite symmetric matrix \mathbf{A} , step size ϵ_g , convergence check value ϵ_0 , parameters ξ and μ ($0 < \xi < \mu < 1$) used in the Armijo criterion and the Wolfe criterion, respectively. Let $c_a = 1$ and set $k = 0$.
- (2) Solve the state determination problem and calculate $f(\mathbf{x}_k)$.
- (3) Solve the adjoint problem with respect to f and calculate $\mathbf{g}(\mathbf{x}_k)$.
- (4) Use Eq. (3.3.6) to calculate \mathbf{y}_g .
- (5) When $k = 0$, let $\mathbf{y}_g = \bar{\mathbf{y}}_g$ and use Eq. (3.3.7) to obtain c_a . Moreover, substitute $\bar{\mathbf{y}}_g/c_a$ into \mathbf{y}_g .
- (6) Let $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{y}_g$. Solve the state determination problem and calculate $f(\mathbf{x}_{k+1})$.

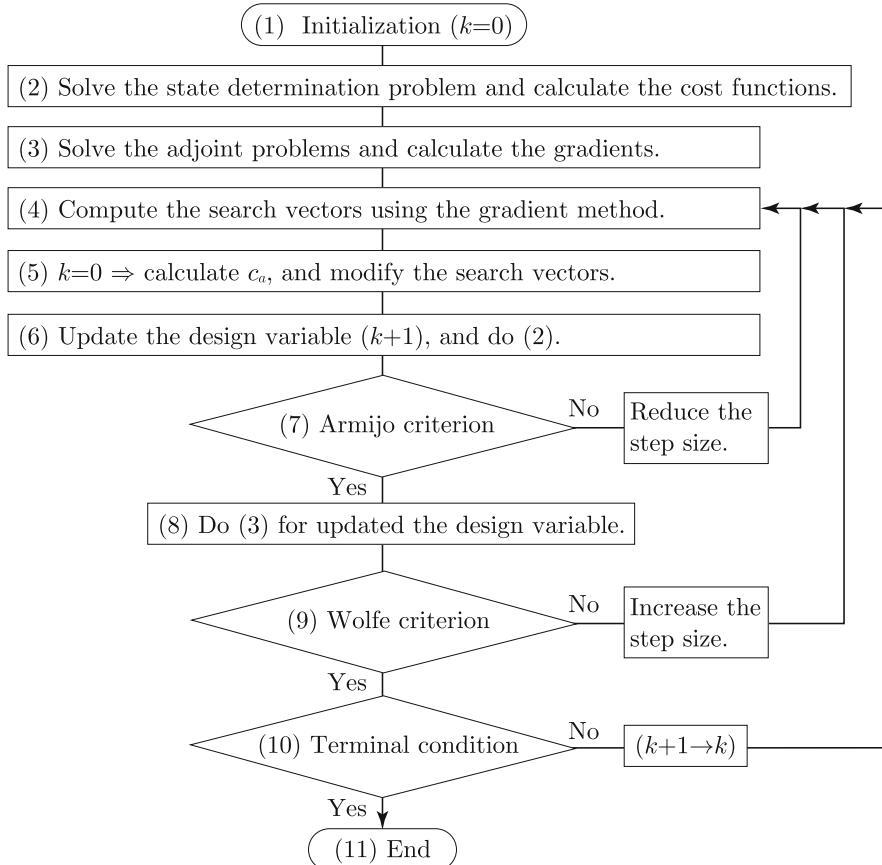


Fig. 3.8 Gradient method algorithm using Armijo criterion and Wolfe criterion

(7) Check the Armijo criterion (Eq. (3.4.3)).

- Proceed to the next step if satisfied.
- Otherwise, suppose $\alpha > 1$, substitute αc_a into c_a , and $\alpha \mathbf{y}_g$ into \mathbf{y}_g , then return to (4).

(8) Calculate $\mathbf{g}(\mathbf{x}_{k+1})$.

(9) Check the Wolfe criterion (Eq. (3.4.6)).

- Proceed to the next step if satisfied.
- Otherwise, suppose $\beta < (0, 1)$, substitute βc_a into c_a and substitute $\beta \mathbf{y}_g$ into \mathbf{y}_g , then return to (4).

(10) Check termination condition $|f_0(\mathbf{x}_{k+1}) - f_0(\mathbf{x}_k)| \leq \epsilon_0$.

- Proceed to (11) when the stopping criterion is satisfied.
- Otherwise, substitute $k + 1$ into k and return to (4).

(11) Complete the calculation. □

Concerning the sequence of iterates obtained through the algorithms in which the step size is restricted so that the Armijo criterion and the Wolfe criterion are satisfied, the following results relating to global convergence can be obtained.

Theorem 3.4.6 (Global Convergence Theorem) *Let $X = \mathbb{R}^d$. Suppose that the function $f : X \rightarrow \mathbb{R}$ has a lower bound, is differentiable in the neighborhood $L = \{\mathbf{x} \in X \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$ of the level set at $\mathbf{x}_0 \in X$ and the gradient \mathbf{g} is Lipschitz continuous (Definition 4.3.1) in L . Let the search vector at \mathbf{x}_k be \mathbf{y}_{gk} and suppose \mathbf{y}_{gk} satisfies $\cos \theta_k > 0$ with respect to descent angle θ_k . Furthermore, suppose that the step size $\|\bar{\epsilon}_g \bar{\mathbf{y}}_g\|_X$ satisfies the Armijo criterion and the Wolfe criterion. Under these assumptions, the sequence of iterates $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$ generated by the gradient method satisfies*

$$\sum_{k \in \mathbb{N}} \|\mathbf{g}(\mathbf{x}_k)\|_{X'}^2 \cos^2 \theta_k < \infty. \quad (3.4.7)$$

□

Proof From the Armijo criterion, we know $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$ is in the neighborhood of L . Moreover, the Wolfe criterion implies that the inequality

$$(\mu - 1) \mathbf{g}(\mathbf{x}_k) \cdot \mathbf{y}_g \leq (\mathbf{g}(\mathbf{x}_{k+1}) - \mathbf{g}(\mathbf{x}_k)) \cdot \mathbf{y}_g$$

holds. Furthermore, since \mathbf{g} is Lipschitz continuous, we have

$$(\mathbf{g}(\mathbf{x}_{k+1}) - \mathbf{g}(\mathbf{x}_k)) \cdot \mathbf{y}_g \leq \beta \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_X \|\mathbf{y}_g\|_X = \bar{\epsilon}_g \beta \|\mathbf{y}_g\|_X^2$$

for some $\beta > 0$. From these equations, we get

$$\bar{\epsilon}_g \geq \frac{(\mathbf{g}(\mathbf{x}_{k+1}) - \mathbf{g}(\mathbf{x}_k)) \cdot \mathbf{y}_g}{\beta \|\mathbf{y}_g\|_X^2} \geq \frac{(\mu - 1) \mathbf{g}(\mathbf{x}_k) \cdot \mathbf{y}_g}{\beta \|\mathbf{y}_g\|_X^2}.$$

Substituting this equation into the Armijo criterion, we obtain

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) + \xi \bar{\epsilon}_g \mathbf{g}(\mathbf{x}_k) \cdot \mathbf{y}_g = f(\mathbf{x}_k) - \xi \frac{\mu - 1}{\beta} \left(\frac{\mathbf{g}(\mathbf{x}_k) \cdot \mathbf{y}_g}{\|\mathbf{y}_g\|_X} \right)^2 \\ &= f(\mathbf{x}_k) - \xi \frac{\mu - 1}{\beta} \|\mathbf{g}(\mathbf{x}_k)\|_{X'}^2 \cos^2 \theta_k. \end{aligned}$$

Therefore, we have

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \xi \frac{\mu - 1}{\beta} \sum_{k \in \{0, \dots, m\}} \|\mathbf{g}(\mathbf{x}_k)\|_{X'}^2 \cos^2 \theta_k.$$

Since f is bounded below, the desired result follows. This proves the theorem. \square

Equation (3.4.7) is called Zoutendijk condition. If the result of Theorem 3.4.6 and the necessary conditions for an infinite series to converge, $\lim_{k \rightarrow \infty} \|\mathbf{g}(\mathbf{x}_k)\|_{X'}^2 \cos^2 \theta_k = 0$, are used, a result such as the one that follows is obtained.

Corollary 3.4.7 (Global Convergence Theorem) *In addition to the suppositions of Theorem 3.4.6, if \mathbf{y}_g is not asymptotic to the direction which crosses $-\mathbf{g}(\mathbf{x}_k)$ orthogonally, i.e., when $\cos \theta_k > 0$, we have that*

$$\lim_{k \rightarrow \infty} \mathbf{g}(\mathbf{x}_k) = \mathbf{0}_{X'}.$$

\square

This result shows that given an appropriate problem setting, if the search direction is obtained via the gradient method and the step size is chosen so that the Armijo criterion and the Wolfe criterion are satisfied, the generated sequence of iterates $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$ has global convergence.

For the rest of this section, we shall introduce the conjugate gradient method as an extension of the gradient method. Let us first define the conjugates between vectors as follows.

Definition 3.4.8 (Conjugate) Suppose $\mathbf{B} \in \mathbb{R}^{d \times d}$ is a positive definite real symmetric matrix. Let $\mathbf{x} \in X$ and $\mathbf{y} \in X$. If $\mathbf{x} \cdot (\mathbf{B}\mathbf{y}) = 0$, then \mathbf{x} and \mathbf{y} are said to be conjugate. \square

In view of Problem 3.4.2, the conjugate gradient method is given as follows.

Problem 3.4.9 (Conjugate Gradient Method) For each $\mathbf{x}_0 \in X$, the search direction $\bar{\mathbf{y}}_{g0}$ and the parameter $\bar{\epsilon}_{g0}$ which adjusts the step size are obtained through the steepest gradient method and the strict line search method, respectively. For each $k \in \mathbb{N}$, provide a value to $\bar{\mathbf{y}}_{gk-1}$ and obtain $\bar{\mathbf{y}}_{gk}$ such that it is conjugate to $\bar{\mathbf{y}}_{gk-1}$. In addition, find the value of the parameter $\bar{\epsilon}_{gk}$ which adjusts the step size based on the strict line search method. \square

Figure 3.9 shows a geometric illustration of the search vector obtained via the conjugate gradient method when $X = \mathbb{R}^2$. By $\bar{\mathbf{y}}_{g0}$ and $\bar{\mathbf{y}}_{g1}$ being chosen so that they are conjugates, making it a two-dimensional vector space Problem 3.4.2, the minimum point can be achieved by seeking the search vector only twice.

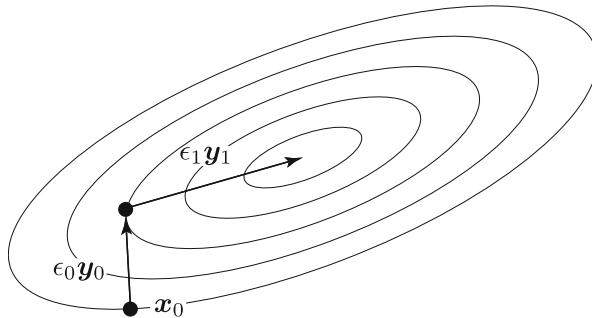


Fig. 3.9 Search vector obtained via conjugate gradient method

Let us show an example of a conjugate gradient method. Let $\mathbf{x}_0 = \mathbf{0}_X$. Use the steepest gradient method to set the search direction to be $\bar{\mathbf{y}}_{g0} = -\mathbf{g}_0 = -\mathbf{g}(\mathbf{x}_0) = -\mathbf{B}\mathbf{x}_0 - \mathbf{b} = -\mathbf{b}$. If, with respect to $k \in \mathbb{N}$, $\bar{\mathbf{y}}_{gk}$ and \mathbf{g}_k are given, seek

$$\bar{\epsilon}_{gk} = \frac{\bar{\mathbf{y}}_{gk} \cdot \mathbf{g}_k}{\bar{\mathbf{y}}_{gk} \cdot (\mathbf{B}\bar{\mathbf{y}}_{gk})} \quad (3.4.8)$$

using Eq. (3.4.2) (the strict line search method). Furthermore, generate a sequence of iterates for $k \in \mathbb{N}$ using

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \bar{\epsilon}_{gk-1} \bar{\mathbf{y}}_{gk-1}, \quad (3.4.9)$$

$$\mathbf{g}_k = \mathbf{g}_{k-1} + \bar{\epsilon}_{gk-1} \mathbf{B} \bar{\mathbf{y}}_{gk-1}, \quad (3.4.10)$$

$$\beta_k = \frac{\mathbf{g}_k \cdot \mathbf{g}_k}{\mathbf{g}_{k-1} \cdot \mathbf{g}_{k-1}}, \quad (3.4.11)$$

$$\bar{\mathbf{y}}_{gk} = -\mathbf{g}_k + \beta_k \bar{\mathbf{y}}_{gk-1}. \quad (3.4.12)$$

In this case, $\bar{\mathbf{y}}_{gk-1}$ and $\bar{\mathbf{y}}_{gk}$ are conjugates (Practice 3.3).

Equation (3.4.11) is called the Fletcher–Reeves formula. Moreover, its equivalent

$$\beta_k = \frac{\mathbf{g}_k \cdot (\mathbf{g}_k - \mathbf{g}_{k-1})}{\mathbf{g}_{k-1} \cdot \mathbf{g}_{k-1}}$$

is called the Polak–Ribière formula. Several formulae other than these are known. These formulae may be equivalent with respect to second-order optimization problems (Problem 3.4.2) but give different results when $\mathbf{g}_k = \mathbf{g}(\mathbf{x}_k)$ is used in non-linear optimization problems which are not second-order.

3.5 Newton Method

In the gradient method, the gradient \mathbf{g} was used to obtain the search direction. The step size was determined in order to satisfy the strict line search method, the Armijo criterion or the Wolfe criterion. In what follows, we shall consider a method for obtaining the search direction and step size simultaneously by using \mathbf{g} and the Hesse matrix \mathbf{H} . This method is called the Newton method. This technique is used to obtain a search vector $\mathbf{y}_g \in X$ by ignoring $\mathcal{O}(\|\mathbf{y}_g\|_X)$ in the Taylor expansion of $\mathbf{g} \cdot \mathbf{y}$ with respect to an arbitrary $\mathbf{y} \in X$ about \mathbf{x}_k , i.e.,

$$\mathbf{g}(\mathbf{x}_k + \mathbf{y}_g) \cdot \mathbf{y} = \mathbf{g}(\mathbf{x}_k) \cdot \mathbf{y} + \mathbf{y}_g \cdot (\mathbf{H}(\mathbf{x}_k) \mathbf{y}) + \mathcal{O}(\|\mathbf{y}_g\|_X)$$

and then letting

$$\mathbf{g}(\mathbf{x}_k + \mathbf{y}_g) \cdot \mathbf{y} = \mathbf{g}(\mathbf{x}_k) \cdot \mathbf{y} + \mathbf{y}_g \cdot (\mathbf{H}(\mathbf{x}_k) \mathbf{y}) = 0.$$

In other words, the Newton method can be formally described as follows.

Problem 3.5.1 (Newton Method) Let $X = \mathbb{R}^d$. Let the gradient and the Hesse matrix of f at $\mathbf{x}_k \in X$ which is not a local minimum point with respect to $f \in C^2(X; \mathbb{R})$ be $\mathbf{g}(\mathbf{x}_k)$ and $\mathbf{H}(\mathbf{x}_k)$, respectively. In this case, obtain $\mathbf{y}_g \in X$ such that

$$\mathbf{y}_g \cdot (\mathbf{H}(\mathbf{x}_k) \mathbf{y}) = -\mathbf{g}(\mathbf{x}_k) \cdot \mathbf{y} \quad (3.5.1)$$

is satisfied for all $\mathbf{y} \in X$. □

The Newton method is the gradient method when the positive definite real symmetric matrix \mathbf{A} used in the gradient method (Definition 3.3.1) is replaced by a Hesse matrix and such that $\bar{\epsilon}_g$ is taken to be unity. The following result can be obtained from the Newton method.

Theorem 3.5.2 (Newton Method) Suppose f is twice differentiable in the neighborhood of the local minimum point \mathbf{x} and the Hesse matrix \mathbf{H} is Lipschitz continuous (Definition 4.3.1). Moreover, $\mathbf{H}(\mathbf{x})$ is assumed to be positive definite. In this case, with a point sufficiently close to a local minimum point taken to be \mathbf{x}_0 , the sequence of iterates $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$ generated by the Newton method is second-order convergent. □

Proof Let the local minimum point be \mathbf{x} . From the fact that the Hesse matrix \mathbf{H} is Lipschitz continuous around \mathbf{x} and $\mathbf{H}(\mathbf{x})$ is regular, a point \mathbf{x}_k sufficiently close to a local minimum point can be selected such that

$$\left\| \mathbf{H}^{-1}(\mathbf{x}) (\mathbf{H}(\mathbf{x}_k) - \mathbf{H}(\mathbf{x})) \right\|_{\mathbb{R}^{d \times d}} \leq \left\| \mathbf{H}^{-1}(\mathbf{x}) \right\|_{\mathbb{R}^{d \times d}} \beta \|\mathbf{x}_k - \mathbf{x}\|_{\mathbb{R}^d} < \frac{1}{2} \quad (3.5.2)$$

is satisfied with respect to some $\beta > 0$, where

$$\left\| \mathbf{H}^{-1}(\mathbf{x}) \right\|_{\mathbb{R}^{d \times d}} = \max_{\mathbf{y} \in \mathbb{R}^d, \|\mathbf{y}\|_{\mathbb{R}^d}=1} \left\| \mathbf{H}^{-1}(\mathbf{x}) \mathbf{y} \right\|_{\mathbb{R}^d}.$$

In view of the above relationships, and using a standard result in Banach perturbation theory (cf. [179], p. 240]), we have that

$$\left\| \mathbf{H}^{-1}(\mathbf{x}_k) \right\|_{\mathbb{R}^{d \times d}} \leq \frac{\left\| \mathbf{H}^{-1}(\mathbf{x}) \right\|_{\mathbb{R}^{d \times d}}}{1 - \left\| \mathbf{H}^{-1}(\mathbf{x}) (\mathbf{H}(\mathbf{x}_k) - \mathbf{H}(\mathbf{x})) \right\|_{\mathbb{R}^{d \times d}}} < 2 \left\| \mathbf{H}^{-1}(\mathbf{x}) \right\|_{\mathbb{R}^{d \times d}}. \quad (3.5.3)$$

Now, using $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{y}_g$, Eq. (3.5.1) and $\mathbf{g}(\mathbf{x}) = \mathbf{0}_{\mathbb{R}^d}$, we obtain

$$\begin{aligned} \mathbf{x}_{k+1} - \mathbf{x} &= \mathbf{x}_k - \mathbf{H}^{-1}(\mathbf{x}_k) \mathbf{g}(\mathbf{x}_k) - \mathbf{x} + \mathbf{H}^{-1}(\mathbf{x}_k) \mathbf{g}(\mathbf{x}) \\ &= -\mathbf{H}^{-1}(\mathbf{x}_k) \{ \mathbf{g}(\mathbf{x}_k) - \mathbf{g}(\mathbf{x}) - \mathbf{H}(\mathbf{x}_k)(\mathbf{x}_k - \mathbf{x}) \}. \end{aligned} \quad (3.5.4)$$

On the other hand, we have

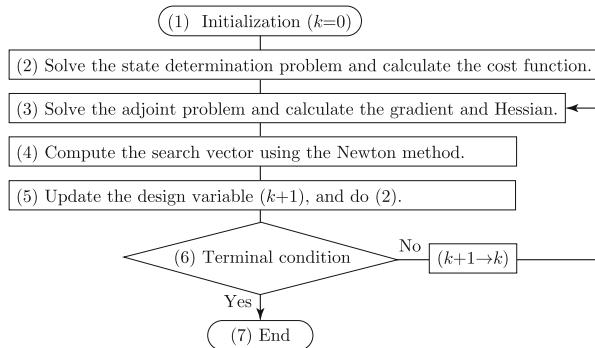
$$\begin{aligned} &\| \mathbf{g}(\mathbf{x}_k) - \mathbf{g}(\mathbf{x}) - \mathbf{H}(\mathbf{x}_k)(\mathbf{x}_k - \mathbf{x}) \|_{\mathbb{R}^d} \\ &= \left\| \int_0^1 (\mathbf{H}(\mathbf{x} + t(\mathbf{x}_k - \mathbf{x})) - \mathbf{H}(\mathbf{x}_k))(\mathbf{x}_k - \mathbf{x}) dt \right\|_{\mathbb{R}^d} \\ &\leq \| \mathbf{x}_k - \mathbf{x} \|_{\mathbb{R}^d} \int_0^1 \| \mathbf{H}(\mathbf{x} + t(\mathbf{x}_k - \mathbf{x})) - \mathbf{H}(\mathbf{x}_k) \|_{\mathbb{R}^{d \times d}} dt \\ &\leq \| \mathbf{x}_k - \mathbf{x} \|_{\mathbb{R}^d} \int_0^1 \beta \| \mathbf{x}_k - \mathbf{x} \|_{\mathbb{R}^d} (1-t) dt \\ &= \frac{1}{2} \beta \| \mathbf{x}_k - \mathbf{x} \|_{\mathbb{R}^d}^2. \end{aligned}$$

Therefore, from Eqs. (3.5.2), (3.5.3) and (3.5.4), it follows that

$$\| \mathbf{x}_{k+1} - \mathbf{x} \|_{\mathbb{R}^d} \leq \frac{1}{2} \left\| \mathbf{H}^{-1}(\mathbf{x}_k) \right\|_{\mathbb{R}^{d \times d}} \beta \| \mathbf{x}_k - \mathbf{x} \|_{\mathbb{R}^d}^2 < \frac{1}{2} \| \mathbf{x}_k - \mathbf{x} \|_{\mathbb{R}^d}. \quad (3.5.5)$$

The relationship between the right-most side and the left-hand side of Eq. (3.5.5) confirms the convergence of the sequence of iterates to the local minimum point \mathbf{x} . Meanwhile, the quadratic convergence of the method follows from the relationship between the first right-hand side and the left-hand side of Eq. (3.5.5). \square

An example of an algorithm using the Newton method is shown below. Figure 3.10 illustrates an overview of the method.

Fig. 3.10 Newton algorithm

Algorithm 3.4 (Newton Method) In Problem 3.1.1, f_0 is written as f and all inequality constraints are taken to be inactive.

- (1) Determine the initial value \mathbf{x}_0 and convergence criterion value ϵ_0 . Set $k = 0$.
- (2) Solve the state determination problem and calculate $f(\mathbf{x}_k)$.
- (3) By solving the adjoint problem with respect to f , calculate $\mathbf{g}(\mathbf{x}_k)$ and $\mathbf{H}(\mathbf{x}_k)$.
- (4) Calculate \mathbf{y}_g using Eq. (3.5.1).
- (5) Let $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{y}_g$. Solve the state determination problem and calculate $f(\mathbf{x}_{k+1})$.
- (6) Check the termination condition $|f_0(\mathbf{x}_{k+1}) - f_0(\mathbf{x}_k)| \leq \epsilon_0$.
 - Proceed to (7) when the termination condition is satisfied.
 - Otherwise, substitute $k + 1$ into k and return to (3).
- (7) Complete the calculation. □

Let us emphasize the following properties of the Newton method.

Remark 3.5.3 (Newton Method) The Newton method has the following quality:

- (1) The Newton method requires the Hesse matrix. The amount of calculation of the Hesse matrix is proportional to the square of the design variable if the matrix is dense. However, when the Hesse matrix is a diagonal matrix, it is proportional to the design variable. Actually, the Hesse matrix in Problem 1.1.4 in Chap. 1 was a diagonal matrix.
- (2) The Newton method has convergence of order two (Theorem 3.5.2).
- (3) If the Hesse matrix is not positive definite, there may be cases when convergence does not occur with the Newton method. Moreover, if it is indefinite, it may converge to the local maximum value (Fig. 3.11).
- (4) When the Hesse matrix \mathbf{H} is a singular (non-invertible) matrix, or there is a large condition number (rate of the maximum eigenvalue to the minimum eigenvalue) of the matrix even if it is not singular, the calculation of the inverse matrix becomes difficult. In fact, if the Hesse matrix is close to a singular matrix, the inverted matrix may be unstable and the solution may diverge. □

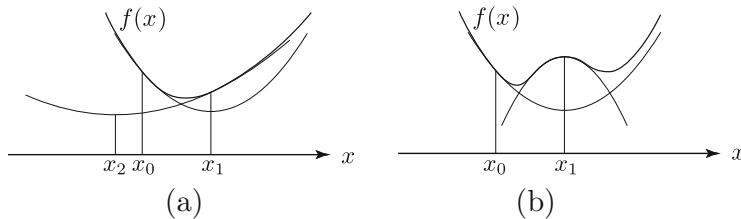


Fig. 3.11 Newton method. (a) Non-convergent case. (b) Convergent case at a local maximizer

Looking at it in this way, the gradient method (Problem 3.3.1) is a method which replaces the Hesse matrix in the Newton method (Problem 3.5.1) with a positive definite real symmetric matrix. Hence, by updating the positive definite symmetric matrices so that they are asymptotic to the Hesse matrix, gradient methods with qualities similar to the Newton method can be studied. These are called quasi-Newton methods. The following are among the known representative updated equations. For details, we refer the interested readers to textbooks on mathematical programming.

- Davidon–Fletcher–Powell method
- Broyden–Fletcher–Goldfarb–Shanno method
- Broyden method

The principle of the Newton method is also used when seeking for solutions of non-linear equations. In such a case, it is also called Newton–Raphson method. The Newton–Raphson method will also be used in Algorithm 3.7 which will be shown later. Hence, let us provide an explanation for it here. Problems to which the Newton–Raphson method applies are such as the one below.

Problem 3.5.4 (Non-linear Equation) Let $X = \mathbb{R}^d$. With respect to $f \in C^1(\mathbb{R}^d; \mathbb{R}^d)$, obtain $x \in X$ such that the equation

$$f(x) = \mathbf{0}_{\mathbb{R}^d} \quad (3.5.6)$$

is satisfied. □

Let the trial point of Problem 3.5.4 be written as x_k for $k \in \mathbb{N}$. Also, assume that f evaluated at x_k and its gradient $\mathbf{G} = (\partial f_i / \partial x_j)_{(i,j) \in \{1, \dots, d\}^2}$ are computable. Under these assumptions, find y_g such that $x_{k+1} = x_k + y_g$ becomes x . The Taylor expansion of f about x_k can be written as

$$f(x_k + y_g) = f(x_k) + \mathbf{G}(x_k) y_g + o\left(\|y_g\|_{\mathbb{R}^d}\right).$$

Here, if $o\left(\|\mathbf{y}_g\|_{\mathbb{R}^d}\right)$ is ignored, then we will get

$$\mathbf{f}(\mathbf{x}_k + \mathbf{y}_g) = \mathbf{f}(\mathbf{x}_k) + \mathbf{G}(\mathbf{x}_k) \mathbf{y}_g = \mathbf{0}_{\mathbb{R}^d}. \quad (3.5.7)$$

Consider the following problem based on Eq. (3.5.7).

Problem 3.5.5 (Newton–Raphson Method) Suppose the function value $\mathbf{f}(\mathbf{x}_k)$ and its gradient $\mathbf{G}(\mathbf{x}_k)$ at the trial point \mathbf{x}_k are given with respect to Problem 3.5.4. Obtain the search vector using

$$\mathbf{y}_g = -\mathbf{G}^{-1}(\mathbf{x}_k) \mathbf{f}(\mathbf{x}_k). \quad (3.5.8)$$

□

The Newton–Raphson method is a method for obtaining a sequence of iterates $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$ which converges to the solution \mathbf{x} of Problem 3.5.4 from $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{y}_g$ using the solution \mathbf{y}_g of Problem 3.5.5. If \mathbf{g} and \mathbf{H} of the Newton method (Problem 3.5.1) are replaced by \mathbf{f} and \mathbf{G} , respectively, it should be understood that it agrees with the Newton–Raphson method (Problem 3.5.5).

Before we end this section, let us consider the Newton method in the case of using the second-order derivative of a cost function derived through the Lagrange multiplier method as explained in the last part of Sect. 1.1.6. Particularly, we assume that the second-order derivative of a cost function is obtained as the Hesse gradient in Eq. (1.1.49). In this case, $\mathbf{H}_0 \mathbf{b}_1$ is obtained as $\mathbf{g}_{\mathbf{H}0}(\mathbf{a}, \mathbf{b}_1)$, then we can consider the gradient method using $\mathbf{g}_{\mathbf{H}0}(\mathbf{a}, \mathbf{b}_1)$ as the gradient. However, in order to obtain $\mathbf{g}_{\mathbf{H}0}(\mathbf{a}, \mathbf{b}_1)$, the adjoint problem (Problem 1.1.6) with respect to the first derivative of the cost function must be solved. Moreover, to solve the adjoint problem, \mathbf{b}_1 should be given. Considering these conditions, we proceed with solving the following problem.

Problem 3.5.6 (Newton Method Using Hesse Gradient) Let $X = \mathbb{R}^d$, $\mathbf{A} \in \mathbb{R}^{d \times d}$ and c_a be a positive definite real symmetric matrix and a positive constant. Moreover, let the gradient, the search vector and the Hesse gradient of f at $\mathbf{x}_k \in X$ which is not a local minimum point with respect to $f \in C^2(X; \mathbb{R})$ be $\mathbf{g}(\mathbf{x}_k)$, $\bar{\mathbf{y}}_g$ and $\mathbf{g}_{\mathbf{H}}(\mathbf{x}_k, \bar{\mathbf{y}}_g)$, respectively. In this case, obtain $\mathbf{y}_g \in X$ such that

$$\mathbf{y}_g \cdot (c_a \mathbf{A}(\mathbf{x}_k) \mathbf{y}) = -(\mathbf{g}(\mathbf{x}_k) + \mathbf{g}_{\mathbf{H}}(\mathbf{x}_k, \bar{\mathbf{y}}_g)) \cdot \mathbf{y} \quad (3.5.9)$$

is satisfied for all $\mathbf{y} \in X$.

□

The solution \mathbf{y}_g of Problem 3.5.6 accords with the solution of the Newton method if $c_a \mathbf{A}(\mathbf{x}_k) = \mathbf{I}$ and $\bar{\mathbf{y}}_g = \mathbf{y}_g$. An example of an algorithm using the solution of Problem 3.5.6 is shown below. Figure 3.12 illustrates an overview of the method.

Algorithm 3.5 (Newton Method Using Hesse Gradient) In Problem 3.1.1, f_0 is written as f and all inequality constraints are taken to be inactive.

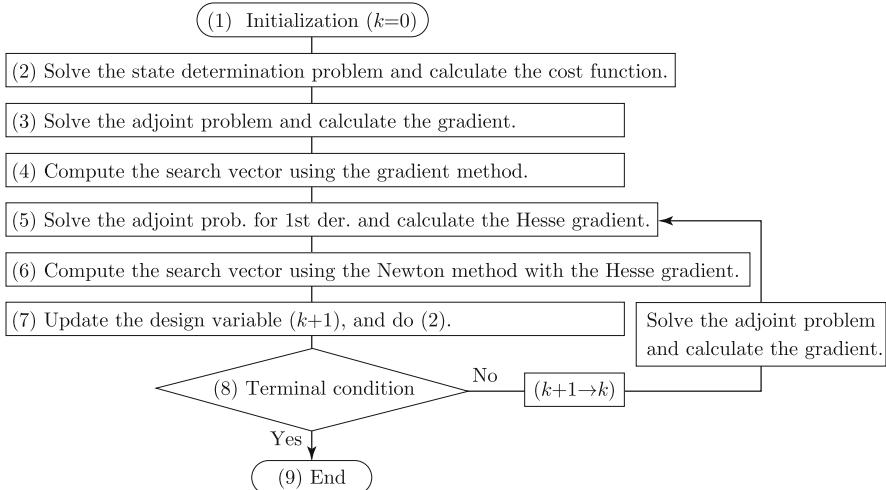


Fig. 3.12 Algorithm of Newton method using Hesse gradient

- (1) Determine the initial value \mathbf{x}_0 and convergence criterion value ϵ_0 . Set $k = 0$.
- (2) Solve the state determination problem and calculate $f(\mathbf{x}_k)$.
- (3) By solving the adjoint problem with respect to f , calculate $\mathbf{g}(\mathbf{x}_k)$.
- (4) Calculate \mathbf{y}_g using the gradient method (Eq. (3.3.6)).
- (5) By solving the adjoint problem with respect to f' , calculate the Hesse gradient $\mathbf{g}_H(\mathbf{x}_k, \mathbf{y}_g)$.
- (6) Calculate \mathbf{y}_g by the Newton method (Eq. (3.5.9)) using $\mathbf{g}_H(\mathbf{x}_k, \mathbf{y}_g)$.
- (7) Let $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{y}_g$. Solve the state determination problem and calculate $f(\mathbf{x}_{k+1})$.
- (8) Check the termination condition $|f_0(\mathbf{x}_{k+1}) - f_0(\mathbf{x}_k)| \leq \epsilon_0$.
 - Proceed to (9) when the termination condition is satisfied.
 - Otherwise, substitute $k + 1$ into k , solve the adjoint problem with respect to f , calculate $\mathbf{g}(\mathbf{x}_k)$ and return to (5).
- (9) Complete the calculation. □

3.6 Augmented Function Methods

Numerical solutions with respect to unconstrained optimization problems were considered from Sects. 3.3 to 3.5. Beyond this section, we shall revert to Problem 3.1.1 and consider the case when the inequality constraint becomes active at the local minimum point. We start this section by considering the method of replacing Problem 3.1.1 with an unconstrained problem by adding constraint functions

multiplied by a constant representing weight to the objective function. Methods such as this are called augmented function methods. Methods for obtaining the minimum value of augmented functions make use of the solutions of unconstrained optimization problems shown from Sects. 3.3 to 3.5.

An augmented function method is a method that sets a point which satisfies all inequality constraints (inner point) as the initial point and utilizes an expansion function designed in a way that it would not fall outside of the admissible set. The method of obtaining a solution to Problem 3.1.1 by using the convergence point of the sequence of iterates $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$ obtained in the following way is called the barrier method or inner point method.

Problem 3.6.1 (Barrier Method, Inner Point Method) Let $\{\rho_k\}_{k \in \mathbb{N}}$ be a positive monotonically decreasing sequence. Suppose that $\mathbf{x}_0 \in X$ is given such that the inequalities $f_1(\mathbf{x}_0) < 0, \dots, f_m(\mathbf{x}_0) < 0$ hold. For $k \in \mathbb{N}$, provide a value for ρ_k and trial point \mathbf{x}_{k-1} and obtain $\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{y}$ such that

$$\min_{\mathbf{y} \in X} \left\{ \hat{f}_k(\mathbf{x}_k, \rho_k) = f_0(\mathbf{x}_k) - \rho_k \sum_{i=1}^m \log(-f_i(\mathbf{x}_k)) \right\}. \quad \square$$

There is another augmented function method which uses an initial point that does not satisfy the inequality constraint conditions and an expansion function such that the trial point is pushed toward being within the admissible set. The method of obtaining a solution for Problem 3.1.1 from the convergence point of the sequence of iterates $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$ obtained in the following way is called the penalty method or outer point method.

Problem 3.6.2 (Penalty Method, Outer Point Method) Let $\{\rho_k\}_{k \in \mathbb{N}}$ be a positive monotonically increasing sequence. Suppose that $\mathbf{x}_0 \in X$ is given. For $k \in \mathbb{N}$, provide a value for ρ_k and trial point \mathbf{x}_{k-1} and obtain $\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{y}$ such that

$$\min_{\mathbf{y} \in X} \left\{ \hat{f}_k(\mathbf{x}_k, \rho_k) = f_0(\mathbf{x}_k) + \rho_k \sum_{i \in \{1, \dots, m\}} \max\{0, f_i(\mathbf{x}_k)\} \right\}. \quad \square$$

From the definitions given above, the augmented function method should be easy to use from the fact that the principles are clear. However, in order to use this method, it is necessary to choose an appropriate monotonically decreasing sequence or monotonically increasing sequence $\{\rho_k\}_{k \in \mathbb{N}}$ depending on the problem. In this book we will not touch upon the details of its selection method.

3.7 Gradient Method for Constrained Problems

In this section, and in Sect. 3.8, we shall turn our attention to a method that employs the KKT conditions. The algorithms that will be shown here will be used in Chap. 7 and beyond. To begin with, let us consider the gradient method for constrained problems.

The admissible set for which the inequality constraints are satisfied with respect to Problem 3.1.1 is written as

$$S = \{x \in X \mid f_1(x) \leq 0, \dots, f_m(x) \leq 0\}. \quad (3.7.1)$$

Moreover, for each $x \in S$, we shall denote by

$$I_A(x) = \{i \in \{1, \dots, m\} \mid f_i(x) \geq 0\} = \{i_1, \dots, i_{|I_A(x)|}\} \quad (3.7.2)$$

the set of subscripts for active constraints. When there is no confusion, $I_A(x_k)$ is written as I_A .

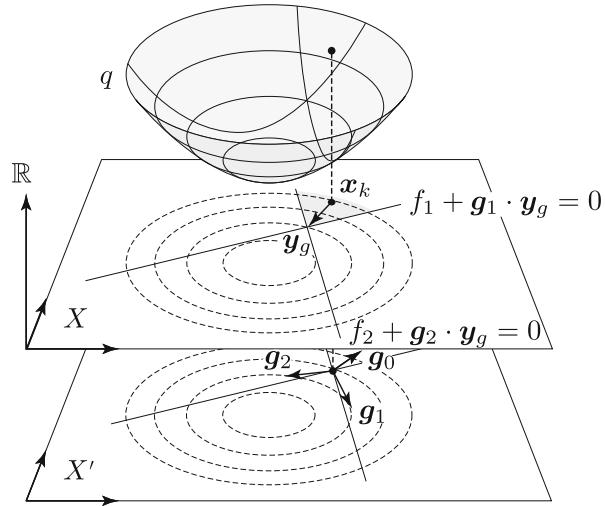
The gradient method was considered as a Newton method when cost function f is approximated by a second-order approximate function $q(y)$ in Eq. (3.3.8) around the trial point $x_k \in S$, for $k \in \mathbb{N}$. So, with respect to Problem 3.1.1, assume the cost function to be $q(y)$ in Eq. (3.3.8) and consider the problem in which the inequality constraint is approximated by a first-order function using the gradient such as the following.

Problem 3.7.1 (Gradient Method for Constrained Problems) For a trial point $x_k \in S$ in Problem 3.1.1, let $f_0(x_k)$, $f_{i_1}(x_k) = 0, \dots, f_{i_{|I_A|}}(x_k) = 0$, and $\mathbf{g}_0(x_k)$, $\mathbf{g}_{i_1}(x_k), \dots, \mathbf{g}_{i_{|I_A|}}(x_k)$ be given. Moreover, let $A \in \mathbb{R}^{d \times d}$ be a positive definite real symmetric matrix and c_a be a positive constant. Obtain $x_{k+1} = x_k + y_g$ which satisfies

$$q(y_g) = \min_{y \in X} \left\{ q(y) = \frac{1}{2} y \cdot (c_a A y) + \mathbf{g}_0(x_k) \cdot y + f_0(x_k) \mid \begin{array}{l} f_i(x_k) + \mathbf{g}_i(x_k) \cdot y \leq 0 \text{ for } i \in I_A(x_k) \end{array} \right\}. \quad \square$$

Problem 3.7.1 can be classified as a second-order optimization problem based on the classification of optimization problems in Sect. 2.2. Furthermore, with the fact that A is a positive definite symmetric real matrix, it is a convex optimization problem. Therefore y_g , which satisfies the KKT conditions with respect to this problem, is the minimum point of Problem 3.7.1 (Fig. 3.13). Let us examine a method for finding y_g .

Fig. 3.13 Gradient method for constrained problems



Let the Lagrange function of Problem 3.7.1 be

$$\mathcal{L}_Q(\mathbf{y}, \boldsymbol{\lambda}_{k+1}) = q(\mathbf{y}) + \sum_{i \in I_A(\mathbf{x}_k)} \lambda_{i,k+1} (f_i(\mathbf{x}_k) + \mathbf{g}_i(\mathbf{x}_k) \cdot \mathbf{y}). \quad (3.7.3)$$

The KKT conditions for Problem 3.7.1's minimum point \mathbf{y}_g are as follows:

$$c_a \mathbf{A} \mathbf{y}_g + \mathbf{g}_0(\mathbf{x}_k) + \sum_{i \in I_A(\mathbf{x}_k)} \lambda_{i,k+1} \mathbf{g}_i(\mathbf{x}_k) = \mathbf{0}_{X'}, \quad (3.7.4)$$

$$f_i(\mathbf{x}_k) + \mathbf{g}_i(\mathbf{x}_k) \cdot \mathbf{y}_g \leq 0 \quad \text{for } i \in I_A(\mathbf{x}_k), \quad (3.7.5)$$

$$\lambda_{i,k+1} (f_i(\mathbf{x}_k) + \mathbf{g}_i(\mathbf{x}_k) \cdot \mathbf{y}_g) = 0 \quad \text{for } i \in I_A(\mathbf{x}_k), \quad (3.7.6)$$

$$\lambda_{i,k+1} \geq 0 \quad \text{for } i \in I_A(\mathbf{x}_k). \quad (3.7.7)$$

If the inequality constraints are assumed to be active with respect to $i \in I_A(\mathbf{x}_k)$, Eqs. (3.7.4) and (3.7.5) become

$$\begin{pmatrix} c_a \mathbf{A} & \mathbf{G}^\top \\ \mathbf{G} & \mathbf{0}_{\mathbb{R}^{|I_A| \times |I_A|}} \end{pmatrix} \begin{pmatrix} \mathbf{y}_g \\ \boldsymbol{\lambda}_{k+1} \end{pmatrix} = - \begin{pmatrix} \mathbf{g}_0 \\ (f_i)_{i \in I_A} \end{pmatrix}, \quad (3.7.8)$$

where

$$\mathbf{G}^\top = \left(\mathbf{g}_{i_1}(\mathbf{x}_k) \cdots \mathbf{g}_{i_{|I_A(\mathbf{x}_k)|}}(\mathbf{x}_k) \right).$$

In this case if $\mathbf{g}_{i_1}, \dots, \mathbf{g}_{i_{|I_A|}}$ are linearly independent, Eq. (3.7.8) is solvable about $(\mathbf{y}_g, \lambda_{k+1})$. With respect to the solutions of these simultaneous linear equations, denote by

$$I_I(\mathbf{x}_k) = \{i \in I_A(\mathbf{x}_k) \mid \lambda_{i,k+1} < 0\} \quad (3.7.9)$$

the set of inactive constraint conditions. If it happens that $I_I(\mathbf{x}_k) \neq \emptyset$, $I_A(\mathbf{x}_k) \setminus I_I(\mathbf{x}_k)$ is replaced with $I_A(\mathbf{x}_k)$ and Eq. (3.7.8) should be solved again. The pair $(\mathbf{y}_g, \lambda_{k+1}) \in X \times \mathbb{R}^{|I_A|}$ obtained in this way satisfies Eq. (3.7.4) to Eq. (3.7.7). The iterative method, leaving only the active constraints for each iteration, is called the active set method [127, Section 10.10.6, p. 447].

On the other hand, a method can be considered in which, instead of solving Eq. (3.7.8) directly, results in using the gradient method with respect to f_i for each $i \in I_A(\mathbf{x}_k)$. This method is described as follows. The functions $\mathbf{g}_0, \mathbf{g}_{i_1}, \dots, \mathbf{g}_{i_{|I_A|}}$ are used and the gradient method is applied individually. In other words, $\mathbf{y}_{g0}, \mathbf{y}_{gi_1}, \dots, \mathbf{y}_{gi_{|I_A|}}$ is sought so that

$$\mathbf{y}_{gi} = -(c_a \mathbf{A})^{-1} \mathbf{g}_i \quad (3.7.10)$$

is satisfied. Here, the Lagrange multiplier $\lambda_{k+1} \in \mathbb{R}^{|I_A|}$ is taken to be an unknown variable and

$$\mathbf{y}_g = \mathbf{y}_g(\lambda_{k+1}) = \mathbf{y}_{g0} + \sum_{i \in I_A(\mathbf{x}_k)} \lambda_{i,k+1} \mathbf{y}_{gi}. \quad (3.7.11)$$

It can be verified that \mathbf{y}_g satisfies the first row of Eq. (3.7.8). On the other hand, the second row of Eq. (3.7.8) becomes

$$\begin{aligned} & \begin{pmatrix} \mathbf{g}_{i_1} \cdot \mathbf{y}_{gi_1} & \cdots & \mathbf{g}_{i_1} \cdot \mathbf{y}_{gi_{|I_A|}} \\ \vdots & \ddots & \vdots \\ \mathbf{g}_{i_{|I_A|}} \cdot \mathbf{y}_{gi_1} & \cdots & \mathbf{g}_{i_{|I_A|}} \cdot \mathbf{y}_{gi_{|I_A|}} \end{pmatrix} \begin{pmatrix} \lambda_{i_1,k+1} \\ \vdots \\ \lambda_{i_{|I_A|},k+1} \end{pmatrix} \\ &= - \begin{pmatrix} f_{i_1} + \mathbf{g}_{i_1} \cdot \mathbf{y}_{g0} \\ \vdots \\ f_{i_{|I_A|}} + \mathbf{g}_{i_{|I_A|}} \cdot \mathbf{y}_{g0} \end{pmatrix}, \end{aligned}$$

which can equivalently be written as

$$(\mathbf{g}_i \cdot \mathbf{y}_{gj})_{(i,j) \in I_A^2} (\lambda_{j,k+1})_{j \in I_A} = - (f_i + \mathbf{g}_i \cdot \mathbf{y}_{g0})_{i \in I_A}. \quad (3.7.12)$$

Again, in this case, if $\mathbf{g}_{i_1}, \dots, \mathbf{g}_{i_{|I_A|}}$ are linearly independent, then λ_{k+1} are uniquely determined by Eq. (3.7.12). The active constraint method is then applied

with respect to the solution of these simultaneous linear equations. In other words, when the set $I_I(\mathbf{x}_k)$ in Eq. (3.7.9) is non-empty, $I_A(\mathbf{x}_k) \setminus I_I(\mathbf{x}_k)$ is replaced by $I_A(\mathbf{x}_k)$ and Eq. (3.7.8) is solved again. The pair $(\mathbf{y}_g, \lambda_{k+1}) \in X \times \mathbb{R}^{|I_A|}$ obtained in this way should satisfy Eq. (3.7.4) to Eq. (3.7.7).

Moreover, in Eq. (3.7.12), if all the values of active constraint functions $f_{i_1}, \dots, f_{i_{|I_A|}}$ are zero, even if all of $\mathbf{y}_{g0}, \mathbf{y}_{gi_1}, \dots, \mathbf{y}_{gi_{|I_A|}}$ are multiplied by an arbitrary constant, λ_{k+1} remains unchanged. This shows that even if the step size $\|\mathbf{y}_g\|_X$ has not been appropriately selected, λ_{k+1} can be obtained. This relationship is used in Sect. 3.7.2 in order to determine c_a such that the initial value of the step size becomes the desired size.

The above discussion is summarized as follows: the gradient method for constrained problems is an iterative method, whereby \mathbf{y}_g is updated by either directly solving for the search vector \mathbf{y}_g and the Lagrange multiplier λ_{k+1} using Eq. (3.7.8), or by solving for \mathbf{y}_{gi} with Eq. (3.7.10) for each $i \in I_A(\mathbf{x}_k)$, using those to obtain λ_{k+1} from Eq. (3.7.12), and substituting them into Eq. (3.7.11) in order to obtain \mathbf{y}_g .

Before showing specific algorithms, let us consider several situations. One is a situation whereby the inequality constraint of Problem 3.1.1 is replaced by equality constraint $f_i(\mathbf{x}) = 0$. In reality, the inequality constraints are treated in the same manner when the equality constraints are active, and so, this situation can always arise. This type of equality constraint can be replaced by two inequality constraints $f_i(\mathbf{x}) \leq 0$ and $-f_i(\mathbf{x}) \leq 0$. However, when these two inequality constraints are non-linear, determining \mathbf{x} so that they are strictly satisfied is generally difficult. Hence, there is a need to determine a positive constant ϵ_i and relax the constraint such as by $|f_i(\mathbf{x})| \leq \epsilon_i$. In algorithms that will be shown later, only inequality constraints are assumed; it may be thought that there is no need to relax the constraints using ϵ_i . However, if inequality constraints are active, they have the same meaning as the equality constraints $f_i(\mathbf{x}) = 0$ and there is a need to relax the constraint using a positive constant ϵ_i .

Additionally, we suppose a situation in which all inequality constraints are satisfied at the initial point \mathbf{x}_0 . If this type of condition is not satisfied, $\mathbf{x}_0 \in S$ which satisfies all the inequality constraints can be found by carrying out the following steps for pre-processing. If they cannot be found, there is a need to review the problem set-up.

- (0) Let the cost function f_0 be zero and \mathbf{g}_0 be equal to the zero vector $\mathbf{0}_{\mathbb{R}^d}$ and then repeat the established steps in the algorithm that will be shown later until all the inequality constraints are satisfied.

3.7.1 Simple Algorithm

With all the things looked at already in mind, let us now examine a simple algorithm in the succeeding discussion. In this section, the parameter c_a for adjusting the step

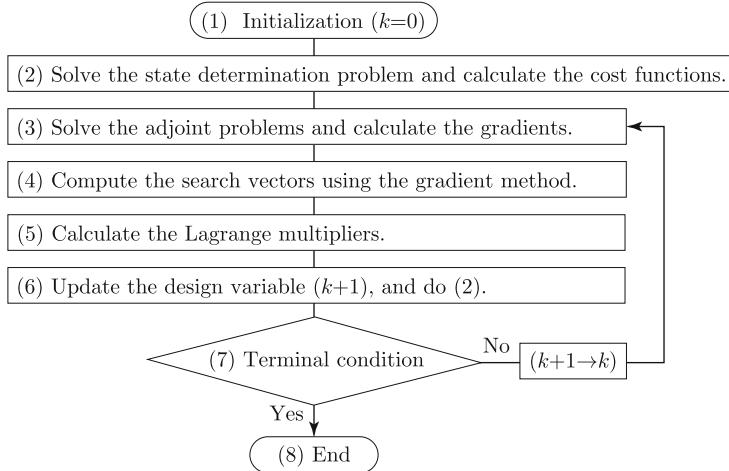


Fig. 3.14 Algorithm of gradient method for constraint problems without parameter adjustment

size is given in advance, and an example of an algorithm when inequality constraint checks are not carried out after updating the design variables is shown. Figure 3.14 shows the flow diagram for the algorithm.

Algorithm 3.6 (Gradient Method Without Parameter Adjustment) Obtain the local minimum point of Problem 3.1.1 in the following way:

- (1) Determine the initial point \mathbf{x}_0 so that the inequality constraints $f_1(\mathbf{x}_0) \leq 0, \dots, f_m(\mathbf{x}_0) \leq 0$ are satisfied. Determine the positive definite matrix \mathbf{A} of Eq. (3.7.10), positive constant c_a for adjusting the step size, positive constant ϵ_0 used for the check of convergence of f_0 as well as the positive constants $\epsilon_1, \dots, \epsilon_m$ providing the permissible ranges of f_1, \dots, f_m . Set $k = 0$.
- (2) Solve the state determination problem for \mathbf{x}_k and calculate $f_0(\mathbf{x}_k), f_1(\mathbf{x}_k), \dots, f_m(\mathbf{x}_k)$. Moreover, let

$$I_{\mathbf{A}}(\mathbf{x}_k) = \{i \in \{1, \dots, m\} \mid f_i(\mathbf{x}_k) \geq -\epsilon_i\}.$$

- (3) Solve the adjoint problem with respect to $f_0, f_{i_1}, \dots, f_{i_{|I_{\mathbf{A}}|}}$ and for \mathbf{x}_k , calculate $\mathbf{g}_0, \mathbf{g}_{i_1}, \dots, \mathbf{g}_{i_{|I_{\mathbf{A}}|}}$.
- (4) Calculate $\mathbf{y}_{g0}, \mathbf{y}_{gi_1}, \dots, \mathbf{y}_{gi_{|I_{\mathbf{A}}|}}$ with Eq. (3.7.10).
- (5) Use Eq. (3.7.12) to seek λ_{k+1} . If $I_{\mathbf{I}}(\mathbf{x}_k)$ in Eq. (3.7.9) is non-empty, replace $I_{\mathbf{A}}(\mathbf{x}_k) \setminus I_{\mathbf{I}}(\mathbf{x}_k)$ by $I_{\mathbf{A}}(\mathbf{x}_k)$ and solve Eq. (3.7.12) again.
- (6) Use Eq. (3.7.11) to seek \mathbf{y}_g , and letting $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{y}_g$, calculate $f_0(\mathbf{x}_{k+1}), f_1(\mathbf{x}_{k+1}), \dots, f_m(\mathbf{x}_{k+1})$. Moreover, define

$$I_{\mathbf{A}}(\mathbf{x}_{k+1}) = \{i \in \{1, \dots, m\} \mid f_i(\mathbf{x}_{k+1}) \geq -\epsilon_i\}.$$

(7) Check the terminal condition $|f_0(\mathbf{x}_{k+1}) - f_0(\mathbf{x}_k)| \leq \epsilon_0$.

- Proceed to (8) when the terminal condition is satisfied.
- Otherwise substitute $k + 1$ into k and revert to (3).

(8) End the calculation. □

Let us seek the trial points with respect to Exercise 1.1.7 (numerical example of mean compliance minimization problem) in Chap. 1 using Algorithm 3.6.

Exercise 3.7.2 (Mean Compliance Minimization Problem) Consider Exercise 1.1.7. Let the initial point be $\mathbf{a}_{(0)} = (1/2, 1/2)^\top$ and use Algorithm 3.6 in order to obtain the trial points for $k \in \{0, 1\}$. Here, the required matrix and numerical values should be determined appropriately. □

Answer The mean compliance $\tilde{f}_0(\mathbf{a})$ and volume constraint function $f_1(\mathbf{a})$ are given by

$$\tilde{f}_0(\mathbf{a}) = \frac{4}{a_1} + \frac{1}{a_2}, \quad (3.7.13)$$

$$f_1(\mathbf{a}) = a_1 + a_2 - 1 \quad (3.7.14)$$

respectively. Moreover, their cross-sectional derivative will be

$$\mathbf{g}_0 = - \begin{pmatrix} \frac{4}{a_1^2} \\ \frac{1}{a_2^2} \end{pmatrix}, \quad (3.7.15)$$

$$\mathbf{g}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \quad (3.7.16)$$

Numerical values are sought along with Algorithm 3.6. Here, the design variable is written as $\mathbf{a}_{(k)}$ for each step number $k \in \mathbb{N}$. The same is true for $\mathbf{b}_{g0(k)}$, $\mathbf{b}_{g1(k)}$ and $\lambda_{1(k)}$.

- (1) At initial point $\mathbf{a}_{(0)} = (1/2, 1/2)^\top$, $f_1(\mathbf{a}_{(0)}) = 0$ is satisfied. Let the positive definite matrix of Eq. (3.7.10) be $\mathbf{A} = \mathbf{I}$, and the positive constant for adjusting the step size be $c_a = 100$ (step size is $\|\mathbf{b}_{g(0)}\|_{\mathbb{R}^2} = 0.0848528$ from calculation shown later on) and $\epsilon_0 = 10^{-3} \tilde{f}_0(\mathbf{a}_{(0)})$, $\epsilon_1 = 10^{-3}$. Set $k = 0$.
- (2) Equations (3.7.13) and (3.7.14) give $\tilde{f}_0(\mathbf{a}_{(0)}) = 10$ and $f_1(\mathbf{a}_{(0)}) = 0$. Moreover, let $I_{\mathbf{A}}(\mathbf{a}_{(0)}) = \{1\}$.
- (3) Equations (3.7.15) and (3.7.16) give $\mathbf{g}_{0(0)} = -(16, 4)^\top$, $\mathbf{g}_{1(0)} = (1, 1)^\top$.
- (4) Equation (3.7.10) gives $\mathbf{b}_{g0(0)} = (0.16, 0.04)^\top$, $\mathbf{b}_{g1(0)} = -(0.01, 0.01)^\top$.
- (5) Equation (3.7.12) gives $\lambda_{1(1)} = 10$.
- (6) Equation (3.7.11) gives $\mathbf{b}_{g(0)} = (0.06, -0.06)^\top$ and letting $\mathbf{a}_{(1)} = \mathbf{a}_{(0)} + \mathbf{b}_{g(0)} = (0.56, 0.44)^\top$ give $\tilde{f}_0(\mathbf{a}_{(1)}) = 9.41558$, $f_1(\mathbf{a}_{(1)}) = 0$. Moreover, let $I_{\mathbf{A}}(\mathbf{a}_{(1)}) = \{1\}$.

- (7) $|\tilde{f}_0(\mathbf{a}_{(1)}) - \tilde{f}_0(\mathbf{a}_{(0)})| = 0.584416 \geq \epsilon_0 = 0.01$ suggests that the terminal condition is not satisfied and hence substitute 1 into k and revert to (3).
- (3) Equations (3.7.15) and (3.7.16) give $\mathbf{g}_{0(1)} = -(12.7551, 5.16529)^\top$, $\mathbf{g}_{1(1)} = (1, 1)^\top$.
- (4) Equation (3.7.10) gives $\mathbf{b}_{g0(1)} = (0.127551, 0.0516529)^\top$ and $\mathbf{b}_{g1(1)} = -(0.01, 0.01)^\top$.
- (5) Equation (3.7.12) gives $\lambda_{1(2)} = 8.9602$.
- (6) Equation (3.7.11) gives $\mathbf{b}_{g(1)} = (0.0379491, -0.0379491)^\top$ and letting $\mathbf{a}_{(2)} = \mathbf{a}_{(1)} + \mathbf{b}_{g(1)} = (0.597949, 0.402051)^\top$ gives $\tilde{f}_0(\mathbf{a}_{(2)}) = 9.17678$, $f_1(\mathbf{a}_{(2)}) = 0$. Moreover, let $I_A(\mathbf{a}_{(1)}) = \{1\}$.
- (7) $|\tilde{f}_0(\mathbf{a}_{(2)}) - \tilde{f}_0(\mathbf{a}_{(1)})| = 0.238804 \geq \epsilon_0 = 0.01$ shows that the terminal condition is not satisfied and so substitute 2 into k and return to (3).

Figure 3.15 illustrates the above computations as well as the later computations. Here, f_0 init denotes the value of f_0 at $k = 0$. \square

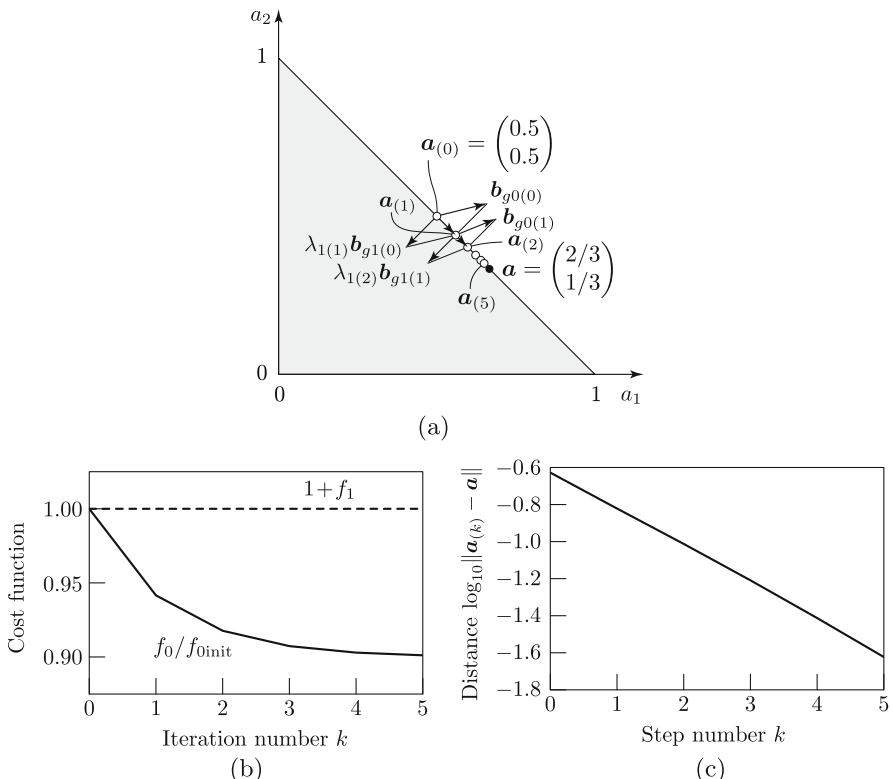


Fig. 3.15 Numerical example of mean compliance minimization problem via gradient method with respect to constraint problem without parameter adjustment. (a) Movement of trial point. (b) History of cost function. (c) Distance $\|\mathbf{a}_k - \mathbf{a}\|_{\mathbb{R}^2}$

Next, let us change the cost function and constraint function of Exercise 1.1.7.

Exercise 3.7.3 (Volume Minimization Problem) Let the cost function and constraint function be

$$f_0(\mathbf{a}) = a_1 + a_2, \quad (3.7.17)$$

$$\tilde{f}_1(\mathbf{a}) = \frac{4}{a_1} + \frac{1}{a_2} - 9, \quad (3.7.18)$$

respectively. In this case, under the constraint which satisfies $\tilde{f}_1(\mathbf{a}) \leq 0$, let the initial point with respect to the problem minimizing $f_0(\mathbf{a})$ (volume minimizing problem with mean compliance constraint) be $\mathbf{a}_{(0)} = (16/31, 4/5)^\top \approx (0.516, 0.8)^\top$ and use Algorithm 3.6 in order to obtain the trial point for $k \in \{0, 1\}$. Here, the required matrices and numerical values should be appropriately determined. \square

Answer The cross-sectional derivatives of the cost functions $f_0(\mathbf{a})$ and $\tilde{f}_1(\mathbf{a})$ are

$$\mathbf{g}_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad (3.7.19)$$

$$\mathbf{g}_1 = - \begin{pmatrix} \frac{4}{a_1^2} \\ \frac{1}{a_2^2} \end{pmatrix}. \quad (3.7.20)$$

The numerical values are sought alongside Algorithm 3.6. In this case, the design variables are again denoted by $\mathbf{a}_{(k)}$ for each step number $k \in \mathbb{N}$. The same applies to $\mathbf{b}_{g0(k)}$, $\mathbf{b}_{g1(k)}$ and $\lambda_{1(k)}$.

- (1) $\tilde{f}_1(\mathbf{a}_{(0)}) = 0$ is satisfied when the initial point is $\mathbf{a}_{(0)} = (16/31, 4/5)^\top$. Let the positive definite matrix of Eq. (3.7.10) be $A = \mathbf{I}$, positive constant $c_a = 10$ which adjusts the step size to be $c_a = 10$ (step size is $\|\mathbf{b}_{g(0)}\|_{\mathbb{R}^2} = 0.089113$ based on calculations shown later) and $\epsilon_0 = 10^{-3}f_0(\mathbf{a}_{(0)})$, $\epsilon_1 = 9 \times 10^{-3}$. Moreover, let $k = 0$.
- (2) Equations (3.7.17) and (3.7.18) give $f_0(\mathbf{a}_{(0)}) = 1.31613$ and $\tilde{f}_1(\mathbf{a}_{(0)}) = 0$. Moreover, let $I_A(\mathbf{a}_{(0)}) = \{1\}$.
- (3) Equations (3.7.15) and (3.7.16) give $\mathbf{g}_{0(0)} = (1, 1)^\top$, $\mathbf{g}_{1(0)} = -(15.0156, 1.5625)^\top$.
- (4) Equation (3.7.10) gives $\mathbf{b}_{g0(0)} = -(0.1, 0.1)^\top$ and $\mathbf{b}_{g1(0)} = (1.50156, 0.15625)^\top$.
- (5) Equation (3.7.12) gives $\lambda_{1(1)} = 0.0727397$.
- (6) Equation (3.7.11) gives $\mathbf{b}_{g(0)} = (0.00922315, -0.0886344)^\top$ and lets $\mathbf{a}_{(1)} = \mathbf{a}_{(0)} + \mathbf{b}_{g(0)} = (0.525352, 0.711366)^\top$ gives $f_0(\mathbf{a}_{(1)}) = 1.23672$, $\tilde{f}_1(\mathbf{a}_{(1)}) = 0.019687$. Moreover, let $I_A(\mathbf{a}_{(1)}) = \{1\}$.

- (7) $|f_0(\mathbf{a}_{(1)}) - f_0(\mathbf{a}_{(0)})| = 0.0794113 \geq \epsilon_0 = 0.00131613$ suggests that the terminal condition is not yet satisfied, so substitute 1 into k and revert to (3).
- (3) Equations (3.7.15) and (3.7.16) give $\mathbf{g}_{0(1)} = (1, 1)^\top$, $\mathbf{g}_{1(1)} = -(14.493, 1.97612)^\top$.
- (4) Equation (3.7.10) gives $\mathbf{b}_{g0(1)} = -(0.1, 0.1)^\top$ and $\mathbf{b}_{g1(1)} = (1.4493, 0.197612)^\top$.
- (5) Equation (3.7.12) gives $\lambda_{1(2)} = 0.0778958$.
- (6) Equation (3.7.11) gives $\mathbf{b}_{g(1)} = (0.0128945, -0.0846068)^\top$ and by letting $\mathbf{a}_{(2)} = \mathbf{a}_{(1)} + \mathbf{b}_{g(1)} = (0.538247, 0.626759)^\top$, $f_0(\mathbf{a}_{(2)}) = 1.16501$ and $\tilde{f}_1(\mathbf{a}_{(2)}) = 0.0270467$ can be obtained. Moreover, let $I_A(\mathbf{a}_{(1)}) = \{1\}$.
- (7) From $|f_0(\mathbf{a}_{(2)}) - f_0(\mathbf{a}_{(1)})| = 0.0717123 \geq \epsilon_0 = 0.00131613$, the terminal condition is seen to not be satisfied, 2 is substituted in for k and reverts to (3).

Figure 3.16 shows these results and later computed values. \square

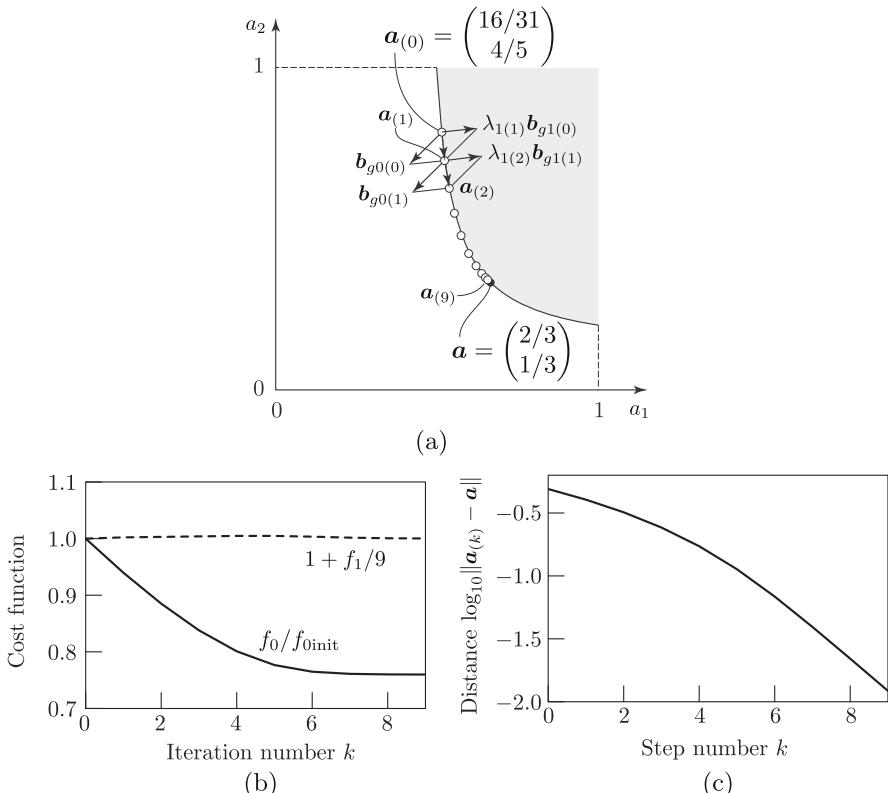


Fig. 3.16 Numerical example of volume minimizing problem using gradient method without parameter adjustment. (a) Movement of trial point. (b) History of cost function. (c) Distance $\|\mathbf{a}_k - \mathbf{a}\|_{\mathbb{R}^2}$

In Exercise 3.7.2, the constraints were always satisfied as $f_1(\mathbf{a}_{(1)}) = f_1(\mathbf{a}_{(2)}) = 0$. However, in Exercise 3.7.3, the constraints were not satisfied since $\tilde{f}_1(\mathbf{a}_{(1)}) = 0.019687$ and $\tilde{f}_1(\mathbf{a}_{(2)}) = 0.0459682$ and their excess values increased with each reiteration. Methods for preventing such a situation will be considered in the next section.

3.7.2 Complicated Algorithm

Let us consider adding the following type of function to a simple algorithm (Algorithm 3.6):

- (i) A function for determining c_a such that, given the initial step size ϵ_g , $\|\mathbf{y}_g\|_X = \epsilon_g$.
- (ii) A function for correcting $\lambda_{k+1} = (\lambda_{i,k+1})_{i \in I_A(\mathbf{x}_{k+1})}$ such that, when design variable is updated to \mathbf{x}_{k+1} , $|f_i(\mathbf{x}_{k+1})| \leq \epsilon_i$ and $\lambda_{i,k+1} \geq 0$ are satisfied with respect to $i \in I_A(\mathbf{x}_{k+1})$.
- (iii) A function for adjusting the permissible values $\epsilon_1, \dots, \epsilon_m$ for constraint functions f_1, \dots, f_m with respect to the convergence check value ϵ_0 for cost function f_0 .
- (iv) A function for adjusting the step size $\|\mathbf{y}_g\|_X$ so that global convergence is guaranteed.

(i) above can be solved, as with Algorithm 3.2, by seeking c_a with Eq. (3.3.7). It is included in step (6) in Algorithm 3.7 which will be shown later.

Moreover, the following types of methods can be thought of with respect to (ii). Even if the value of λ_{k+1} calculated in step (5) of Algorithm 3.6 satisfies the KKT conditions of the gradient method for constrained problems (Problem 3.7.1), the non-linearity of active inequality constraint functions suggest that it is not necessarily satisfied in the specified permissible range at \mathbf{x}_{k+1} .

In order for it to be satisfied within the permissible range, where each of the inequality constraint is specified by \mathbf{x}_{k+1} , λ_{k+1} needs to be amended and this requires modifying Eq. (3.7.11), so that $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{y}_g$ will be updated. In this case, let us consider setting $\lambda_{k+1} = \lambda_{k+10}$, we provide λ_{k+1l} for $l \in \{0, 1, 2, \dots\}$ and repeat the calculations seeking λ_{k+1l+1} . To do this, the Newton–Raphson method (Problem 3.5.5), which is used to solve non-linear equations, will be applied.

The method is described as follows. For each $i \in I_A(\mathbf{x}_{k+1})$, write $f_i(\mathbf{x}_k + \mathbf{y}_g(\lambda_{k+1l}))$ as $\tilde{f}_i(\lambda_{k+1l})$. In the explanation of the Newton–Raphson method, a function $f(\mathbf{x}_k + \mathbf{y}_g) = \mathbf{0}_{\mathbb{R}^d}$ of Eq. (3.5.7) with respect to $k \in \mathbb{N}$ was considered. Here, we shall consider $(\tilde{f}_i(\lambda_{k+1l} + \delta\lambda))_{i \in I_A} = \mathbf{0}_{\mathbb{R}^{|I_A|}}$ where $l \in \{0, 1, 2, \dots\}$. Moreover, by taking into account the fact that $\mathbf{y}_g(\lambda_{k+1l})$ is a first-order function of λ_{k+1l} defined by Eq. (3.7.11), consider

$(\mathbf{g}_i(\lambda_{k+1l}) \cdot \mathbf{y}_{gj}(\lambda_{k+1l}))_{(i,j) \in I_A^2}$ instead of $\mathbf{G}(\mathbf{x}_k)$ of Eq. (3.5.7). In this case, the following can be obtained analogous to Eq. (3.5.8):

$$\begin{aligned}\delta\lambda &= (\delta\lambda_j)_{j \in I_A} \\ &= -(\mathbf{g}_i(\lambda_{k+1l}) \cdot \mathbf{y}_{gj}(\lambda_{k+1l}))_{(i,j) \in I_A^2}^{-1} (f_i(\lambda_{k+1l}))_{i \in I_A}.\end{aligned}\quad (3.7.21)$$

Using $\delta\lambda$ of Eq. (3.7.21), the value of λ_{k+1} is updated using the recursion $\lambda_{k+1l+1} = \lambda_{k+1l} + \delta\lambda$. Furthermore, from Eq. (3.7.11), \mathbf{y}_g is modified, replacing it by $\mathbf{y}_g(\lambda_{k+1l+1})$. As a result, $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{y}_g$ is changed to $\mathbf{x}_{k+1l+1} = \mathbf{x}_k + \mathbf{y}_g(\lambda_{k+1l+1})$. This update is used in step (11) of Algorithm 3.7 which will be shown later.

Now, let us discuss a method to correctly modify λ_{k+1} with respect to Exercise 3.7.3.

Exercise 3.7.4 (Volume Minimizing Problem) Recall from Exercise 3.7.3 that $\tilde{f}_1(\mathbf{a}_{(1)}) = 0.019687$. Using Eq. (3.7.21), correctly adjust $\lambda_{1(1)} = 0.0727397$ and obtain a trial point $\mathbf{a}_{(1)[l]}$ such that $\tilde{f}_1(\mathbf{a}_{(1)[l]}) \leq 10^{-4}$ holds. In this calculation, the adjusted value of λ_1 , for each step number k , should be written as $\lambda_{1(k+1)[l]}$, where l denotes the step number for the adjustment procedure. \square

Answer Consider Eq. (3.7.21) as the example problem. Applying this to Exercise 3.7.3, we get

$$\delta\lambda_1 = -\frac{\tilde{f}_1(\mathbf{a}_{(1)[l]})}{\mathbf{g}_{1(0)[l]} \cdot \mathbf{b}_{g1(0)[l]}}.\quad (3.7.22)$$

When $l = 0$, $\mathbf{a}_{(1)[0]} = \mathbf{a}_{(1)}$, $\mathbf{g}_{1(0)[0]} = \mathbf{g}_{1(0)}$ and $\mathbf{b}_{g1(0)[0]} = \mathbf{b}_{g1(0)}$, giving us $\delta\lambda_1 = 0.000863807$. At this point, λ_1 should be updated to

$$\lambda_{1(1)[1]} = \lambda_1 + \delta\lambda_1 = 0.0736035.$$

If this $\lambda_{1(1)[1]}$ is brought in and Eq. (3.7.11) is used to calculate $\mathbf{b}_{g(0)}$, we will get

$$\mathbf{b}_{g(0)[1]} = (0.0105202, -0.0884995)^\top.$$

Using this search vector to update the design variables will yield

$$\mathbf{a}_{(1)[1]} = \mathbf{a}_{(1)} + \mathbf{b}_{g(0)[1]} = (0.526649, 0.711501)^\top.$$

These calculations imply that $\tilde{f}_1(\mathbf{a}_{(1)[1]}) = 0.00066837 > 10^{-4}$. Clearly, the permitted constraint is not satisfied. Then, let $l = 1$ and repeat the steps above. From Eq. (3.7.22), one obtains

$$\delta\lambda_1 = -\frac{\tilde{f}_1(\mathbf{a}_{(1)[1]})}{\mathbf{g}_{1(0)[1]} \cdot \mathbf{b}_{g1(0)[1]}} = 0.000029326.$$

Here, λ_1 is updated to

$$\lambda_{1(1)[2]} = \lambda_{1(1)[1]} + \delta\lambda_1 = 0.0736328.$$

If $\lambda_{1(1)[2]}$ and Eq. (3.7.11) is used to recalculate $\mathbf{b}_{g(0)}$, then we get

$$\mathbf{b}_{g(0)[2]} = (0.0105202, -0.0884995)^\top.$$

If this search vector is used to update the design variables, then we will have

$$\mathbf{a}_{(1)[2]} = \mathbf{a}_{(1)} + \mathbf{b}_{g(0)[2]} = (0.526693, 0.711505)^\top.$$

At this point, $\tilde{f}_1(\mathbf{a}_{(1)[2]}) = 0.0000243138 \leq 10^{-4}$. \square

Meanwhile, the following method can be used to address the issue stated in (iii). The Lagrange function of the original problem (Problem 3.1.1) can be defined by

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f_0(\mathbf{x}) + \sum_{i \in I_A(\mathbf{x})} \lambda_i f_i(\mathbf{x}). \quad (3.7.23)$$

If with respect to $i \in I_A(\mathbf{x}_k)$, $|f_i(\mathbf{x}_k)| \leq \epsilon_i$ is satisfied in order for $\mathcal{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k) \approx f_0(\mathbf{x}_k)$ to hold, then the inequality

$$\epsilon_0 \gg \sum_{i \in I_A(\mathbf{x}_k)} \lambda_{ki} \epsilon_i \quad (3.7.24)$$

must hold true. Hence, in order for this condition to hold, we require that a positive constant σ be sufficiently small relative to the unity, and that the criteria of constraint permissible values satisfy the following relation:

$$\epsilon_i \leq \frac{\sigma \epsilon_0}{|I_A(\mathbf{x}_k)| \lambda_{ki}}, \quad (3.7.25)$$

for all $i \in I_A(\mathbf{x}_k)$. If there is a case when there is an index i for which this condition does not hold, the criteria can be satisfied by substituting in a value smaller than $\sigma \epsilon_0 / (|I_A(\mathbf{x}_k)| \lambda_{ki})$ in ϵ_i . Such criteria for constraint concerning permissible values are used in Step (12) of Algorithm 3.7 that will be shown later.

On the other hand, a method to determine the step size $\|\mathbf{y}_g\|_X$ (in other words, c_a) so that the Armijo and Wolfe criteria are satisfied with respect to the Lagrange function can be thought of in connection with (iv) above. Theorem 3.4.6 became the basis to guarantee global convergence for unconstrained problems. Here it is assumed that the KKT conditions for the Lagrange multipliers and inequality constraints are satisfied at \mathbf{x}_k and $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{y}_g$ from (ii) and (iii) above (in Algorithm 3.7, which will be shown later, the KKT conditions for the Lagrange multipliers and inequality constraints are satisfied after the step size is adjusted).

Here, the Lagrange function matches f_0 and it becomes possible to use Armijo and Wolfe criteria with respect to unconstrained problems. Let us describe this more formally as follows. Let the Lagrange function of the original problem (Program 3.1.1) be Eq. (3.7.23). Let the gradient of $f_0(\mathbf{x}_k), f_1(\mathbf{x}_k), \dots, f_{|I_A|}(\mathbf{x}_k)$ be written as $\mathbf{g}_0(\mathbf{x}_k), \mathbf{g}_1(\mathbf{x}_k), \dots, \mathbf{g}_{|I_A|}(\mathbf{x}_k)$, respectively, and the gradient of $f_0(\mathbf{x}_k + \mathbf{y}_g), f_1(\mathbf{x}_k + \mathbf{y}_g), \dots, f_{|I_A|}(\mathbf{x}_k + \mathbf{y}_g)$ as $\mathbf{g}_0(\mathbf{x}_k + \mathbf{y}_g), \mathbf{g}_1(\mathbf{x}_k + \mathbf{y}_g), \dots, \mathbf{g}_{|I_A|}(\mathbf{x}_k + \mathbf{y}_g)$ as well. Here, the Armijo criterion with respect to $\xi \in (0, 1)$ is given by

$$\begin{aligned} & \mathcal{L}(\mathbf{x}_k + \mathbf{y}_g, \lambda_{k+1}) - \mathcal{L}(\mathbf{x}_k, \lambda_k) \\ & \leq \xi \left(\mathbf{g}_0(\mathbf{x}_k) + \sum_{i \in I_A(\mathbf{x}_k)} \lambda_{ki} \mathbf{g}_i(\mathbf{x}_k) \right) \cdot \mathbf{y}_g. \end{aligned} \quad (3.7.26)$$

Moreover, the Wolfe criterion with respect to μ ($0 < \xi < \mu < 1$) is given by

$$\begin{aligned} & \mu \left(\mathbf{g}_0(\mathbf{x}_k) + \sum_{i \in I_A(\mathbf{x}_k)} \lambda_{ki} \mathbf{g}_i(\mathbf{x}_k) \right) \cdot \mathbf{y}_g \\ & \leq \left(\mathbf{g}_0(\mathbf{x}_k + \mathbf{y}_g) + \sum_{i \in I_A(\mathbf{x}_{k+1})} \lambda_{i,k+1} \mathbf{g}_i(\mathbf{x}_k + \mathbf{y}_g) \right) \cdot \mathbf{y}_g. \end{aligned} \quad (3.7.27)$$

These criteria are used in steps (8) and (10) of Algorithm 3.7 which is shown below.

An example of an algorithm including the method such as the one above is shown next. Figure 3.17 shows its flow diagram.

Algorithm 3.7 (Gradient Method with Parameter Adjustment) Obtain the local minimum point of Problem 3.1.1 in the following way:

- (1) Determine the initial point \mathbf{x}_0 so that $f_1(\mathbf{x}_0) \leq 0, \dots, f_m(\mathbf{x}_0) \leq 0$ are satisfied. Also, determine the positive definite matrix \mathbf{A} of Eq. (3.7.10) and initial step size ϵ_g , convergence check value ϵ_0 for f_0 , initial values $\epsilon_1, \dots, \epsilon_m$ of permissible ranges for f_1, \dots, f_m , and Armijo and Wolfe standard values ξ and μ ($0 < \xi < \mu < 1$) as well as the standard value σ ($\ll 1$) of the constraint permissible values. Moreover, let $c_a = 1$, and set $k = l = 0$.
- (2) Solve the state determination problem and calculate $f_0(\mathbf{x}_k), f_1(\mathbf{x}_k), \dots, f_m(\mathbf{x}_k)$. Moreover, define

$$I_A(\mathbf{x}_k) = \{i \in \{1, \dots, m\} \mid f_i(\mathbf{x}_k) \geq -\epsilon_i\}.$$

- (3) Solve the adjoint problem with respect to $f_0, f_1, \dots, f_{|I_A|}$ and calculate $\mathbf{g}_0, \mathbf{g}_1, \dots, \mathbf{g}_{|I_A|}$ at \mathbf{x}_k .

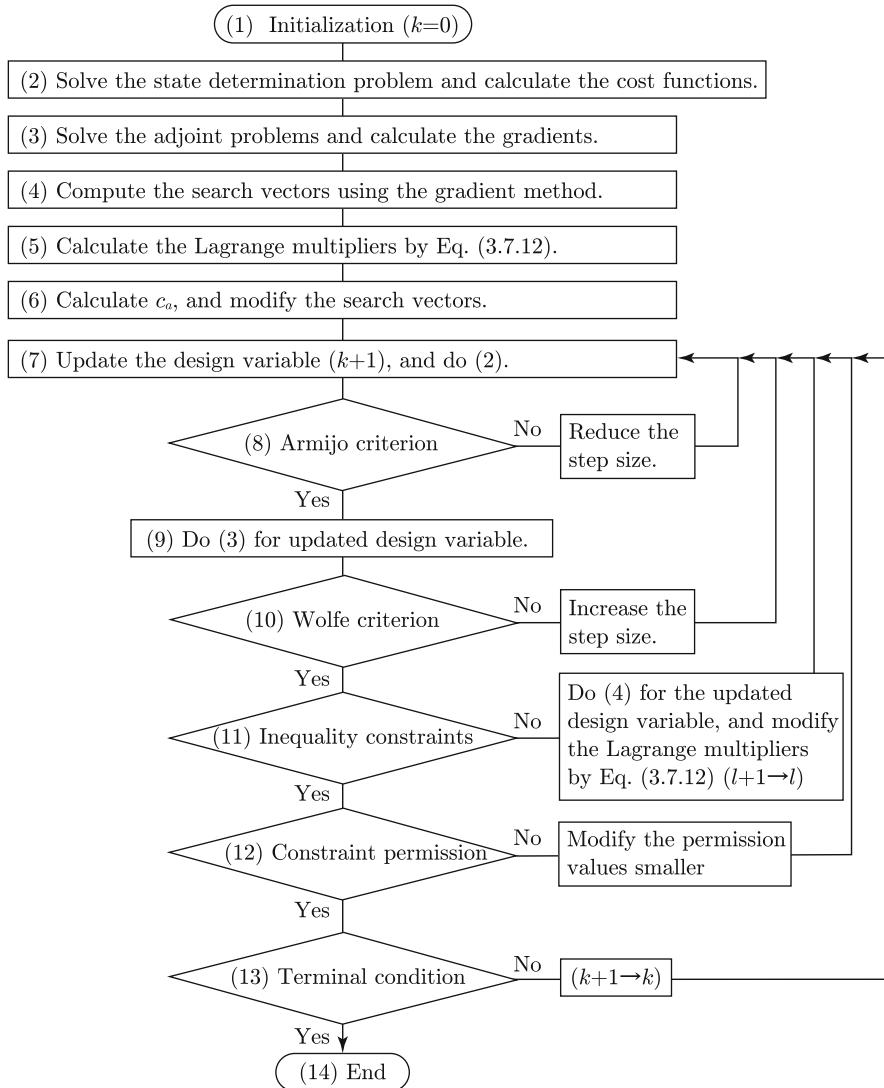


Fig. 3.17 Algorithm for the gradient method with respect to constraint problems with parameter adjustment

- (4) Calculate $\mathbf{y}_{g0}, \mathbf{y}_{gi_1}, \dots, \mathbf{y}_{gi_{|I_A|}}$ using Eq. (3.7.10).
- (5) Seek $\lambda_{k+1} = \lambda_{k+1l}$ using Eq. (3.7.12). If $I_l(\mathbf{x}_k)$ in Eq. (3.7.9) is non-empty, replace $I_A(\mathbf{x}_k) \setminus I_l(\mathbf{x}_k)$ with $I_A(\mathbf{x}_k)$ and solve Eq. (3.7.12) again.
- (6) Obtain \mathbf{y}_g with Eq. (3.7.11). Let $\mathbf{y}_g = \bar{\mathbf{y}}_g$ and use Eq. (3.3.7) to obtain c_a . Moreover, for $i \in I_A(\mathbf{x}_k)$, substitute $\bar{\mathbf{y}}_{gi}/c_a$ into \mathbf{y}_{gi} .

- (7) Let $\mathbf{x}_{k+1l} = \mathbf{x}_k + \mathbf{y}_g(\lambda_{k+1l})$ and calculate $f_0(\mathbf{x}_{k+1l}), f_1(\mathbf{x}_{k+1l}), \dots, f_m(\mathbf{x}_{k+1l})$. Moreover, define

$$I_A(\mathbf{x}_{k+1}) = \{i \in \{1, \dots, m\} \mid f_i(\mathbf{x}_{k+1l}) \geq -\epsilon_i\}.$$

- (8) Let $\lambda_{k+1} = \lambda_{k+1l}$ and check the Armijo criterion (Eq. (3.7.26)).
- If satisfied, proceed to the next step.
 - Otherwise, let $\alpha > 1$, substitute αc_a into c_a and $\mathbf{y}_{g0}/c_a, \mathbf{y}_{gi_1}/c_a, \dots, \mathbf{y}_{gi_{|I_A|}}/c_a$ into $\mathbf{y}_{g0}, \mathbf{y}_{gi_1}, \dots, \mathbf{y}_{gi_{|I_A|}}$ and then revert to (7).
- (9) Calculate $\mathbf{g}_0, \mathbf{g}_{i_1}, \dots, \mathbf{g}_{i_{|I_A|}}$ at \mathbf{x}_{k+1} .
- (10) Let $\lambda_{k+1} = \lambda_{k+1l}$ and check the Wolfe criterion (Eq. (3.7.27)).
- If satisfied, proceed to the next step.
 - Otherwise, let $\beta \in (0, 1)$ and substitute βc_a into c_a and $\beta \mathbf{y}_{g0}, \beta \mathbf{y}_{gi_1}, \dots, \beta \mathbf{y}_{gi_{|I_A|}}$ into $\mathbf{y}_{g0}, \mathbf{y}_{gi_1}, \dots, \mathbf{y}_{gi_{|I_A|}}$ and then return to (7).
- (11) For $i \in I_A(\mathbf{x}_{k+1})$, determine $|f_i(\mathbf{x}_{k+1})| \leq \epsilon_i$.
- If satisfied, proceed to the next step.
 - Otherwise, at \mathbf{x}_{k+1} , calculate $\mathbf{g}_0, \mathbf{g}_{i_1}, \dots, \mathbf{g}_{i_{|I_A|}}$ and also $\mathbf{y}_{g0}, \mathbf{y}_{gi_1}, \dots, \mathbf{y}_{gi_{|I_A|}}$ using Eq. (3.7.10), seek $\delta\lambda$ with Eq. (3.7.21) and let $\lambda_{k+1l+1} = \lambda_{k+1l} + \delta\lambda$. Afterwards, substitute $l + 1$ into l and return to (7).
- (12) For $i \in I_A(\mathbf{x}_{k+1})$, check the criteria for the permissible values of the constraint (Eq. (3.7.25)).
- If satisfied, proceed to the next step.
 - Otherwise, let $\beta \in (0, 1)$ with respect to unsatisfied i , substitute $\beta\sigma\epsilon_0 / (|I_A(\mathbf{x}_{k+1})| \lambda_{k+1l})$ into ϵ_i and revert to (7).
- (13) Check the terminal condition $|f_0(\mathbf{x}_{k+1}) - f_0(\mathbf{x}_k)| \leq \epsilon_0$.
- When the terminal condition is satisfied, proceed to the next step.
 - Otherwise, substitute $k + 1$ into k , let $l = 0$ and revert to (7).
- (14) End the calculation. □

3.8 Newton Method for Constrained Problems

If Hesse matrices of cost functions relating to the variation of \mathbf{x} can be obtained, the Newton method can be used instead of the gradient method. We shall refer to this method as the Newton method for constrained problems. In this case, the Hesse matrices of f_0, f_1, \dots, f_m are expressed as $\mathbf{H}_0, \mathbf{H}_1, \dots, \mathbf{H}_m$, respectively. Moreover, if there is no confusion, $I_A(\mathbf{x}_k)$ is written as I_A .

The Lagrange function \mathcal{L}_S of Eq. (3.7.3) defined with respect to the gradient method for constrained problems (Problem 3.7.1) is then replaced by

$$\begin{aligned} \mathcal{L}_Q(\mathbf{y}, \boldsymbol{\lambda}_{k+1}) &= \frac{1}{2} \mathbf{y} \cdot (\mathbf{H}_0(\mathbf{x}_k) \mathbf{y}) + \mathbf{g}_0(\mathbf{x}_k) \cdot \mathbf{y} + f_0(\mathbf{x}_k) \\ &+ \sum_{i \in I_A(\mathbf{x}_k)} \left\{ \lambda_{i,k+1} (f_i + \mathbf{g}_i(\mathbf{x}_k) \cdot \mathbf{y}) + \lambda_{i,k} \frac{1}{2} \mathbf{y} \cdot (\mathbf{H}_i(\mathbf{x}_k) \mathbf{y}) \right\}. \end{aligned} \quad (3.8.1)$$

In the above formula, $\lambda_k = (\lambda_{i,k})_i$ denotes the Lagrange multiplier obtained in the previous step. When $k = 0$, it is supposed that it is determined via a method used in the gradient method for constrained problems. Let \mathcal{L}_S denote the Lagrange function in the next problem.

Problem 3.8.1 (Newton Method for Constrained Problems) Let $\mathbf{x}_k \in X$ be a trial point in Problem 3.1.1, and $\lambda_k \in \mathbb{R}^{|I_A|}$ be the Lagrange multiplier satisfying Eq. (3.7.5) to Eq. (3.7.7). Moreover, let $f_0(\mathbf{x}_k)$, $f_{i_1}(\mathbf{x}_k) = 0, \dots, f_{i_{|I_A|}}(\mathbf{x}_k) = 0$ as well as $\mathbf{g}_0(\mathbf{x}_k)$, $\mathbf{g}_{i_1}(\mathbf{x}_k), \dots, \mathbf{g}_{i_{|I_A|}}(\mathbf{x}_k)$ and $\mathbf{H}_0(\mathbf{x}_k)$, $\mathbf{H}_{i_1}(\mathbf{x}_k), \dots, \mathbf{H}_{i_{|I_A|}}(\mathbf{x}_k)$ be known, and define

$$\mathbf{H}_{\mathcal{L}}(\mathbf{x}_k) = \mathbf{H}_0(\mathbf{x}_k) + \sum_{i \in I_A(\mathbf{x}_k)} \lambda_{i,k} \mathbf{H}_i(\mathbf{x}_k). \quad (3.8.2)$$

Under these assumptions, find $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{y}_g$ which satisfies

$$q(\mathbf{y}_g) = \min_{\mathbf{y} \in X} \left\{ q(\mathbf{y}) = \frac{1}{2} \mathbf{y} \cdot (\mathbf{H}_{\mathcal{L}}(\mathbf{x}_k) \mathbf{y}) + \mathbf{g}_0(\mathbf{x}_k) \cdot \mathbf{y} + f_0(\mathbf{x}_k) \mid \right. \\ \left. f_i(\mathbf{x}_k) + \mathbf{g}_i(\mathbf{x}_k) \cdot \mathbf{y} \leq 0 \text{ for } i \in I_A(\mathbf{x}_k) \right\}. \quad \square$$

Problem 3.8.1 can be classified as a second-order optimization problem. The expression $\mathbf{H}_{\mathcal{L}}(\mathbf{x}_k)$ need not be a positive definite real matrix, but if it is, Problem 3.8.1 is a convex optimization problem. Let us also consider the method for finding \mathbf{y}_g using KKT conditions with respect to this problem.

The KKT conditions at minimum point \mathbf{y}_g of Problem 3.8.1 are as follows:

$$\mathbf{H}_{\mathcal{L}}(\mathbf{x}_k) \mathbf{y}_g + \mathbf{g}_0(\mathbf{x}_k) + \sum_{i \in I_A(\mathbf{x}_k)} \lambda_{i,k+1} \mathbf{g}_i(\mathbf{x}_k) = \mathbf{0}_{X'}, \quad (3.8.3)$$

$$f_i(\mathbf{x}_{k+1}) = f_i(\mathbf{x}_k) + \mathbf{g}_i(\mathbf{x}_k) \cdot \mathbf{y}_g \leq 0 \quad \text{for } i \in I_A(\mathbf{x}_k), \quad (3.8.4)$$

$$\lambda_{i,k+1} (f_i(\mathbf{x}_k) + \mathbf{g}_i(\mathbf{x}_k) \cdot \mathbf{y}_g) = 0 \quad \text{for } i \in I_A(\mathbf{x}_k), \quad (3.8.5)$$

$$\lambda_{i,k+1} \geq 0 \quad \text{for } i \in I_A(\mathbf{x}_k). \quad (3.8.6)$$

The pair $(\mathbf{y}_g, \lambda_{k+1}) \in X \times \mathbb{R}^{|I_A|}$ satisfying these conditions can be obtained in the following way. Suppose that those inequality constraints are active for all $i \in I_A(\mathbf{x}_k)$. Then, Eqs. (3.8.3) and (3.8.4) become

$$\begin{pmatrix} \mathbf{H}_{\mathcal{L}} & \mathbf{G}^T \\ \mathbf{G} & \mathbf{0}_{\mathbb{R}^{|I_A|} \times |I_A|} \end{pmatrix} \begin{pmatrix} \mathbf{y}_g \\ \lambda_{k+1} \end{pmatrix} = - \begin{pmatrix} \mathbf{g}_0 \\ (f_i)_{i \in I_A} \end{pmatrix}, \quad (3.8.7)$$

where

$$\mathbf{G}^T = (\mathbf{g}_{i_1}, \dots, \mathbf{g}_{i_{|I_A|}}).$$

If $\mathbf{g}_{i_1}, \dots, \mathbf{g}_{i_{|I_A|}}$ are linearly independent and $\mathbf{H}_{\mathcal{L}}$ is regular, Eq. (3.8.7) becomes solvable around $(\mathbf{y}_g, \lambda_{k+1})$. With respect to these simultaneous first-order equations, define

$$I_I(\mathbf{x}_k) = \{i \in I_A(\mathbf{x}_k) \mid \lambda_{i,k+1} < 0\} \quad (3.8.8)$$

as the set of inactive constraint conditions, and when $I_I(\mathbf{x}_k)$ is non-empty, replace $I_A(\mathbf{x}_k) \setminus I_I(\mathbf{x}_k)$ by $I_A(\mathbf{x}_k)$ and then solve Eq. (3.8.7) again. The pair $(\mathbf{y}_g, \lambda_{k+1}) \in X \times \mathbb{R}^{|I_A|}$ obtained in this way satisfies Eq. (3.8.3) to Eq. (3.8.6).

Moreover, as also seen in the gradient method for constrained problems (Sect. 3.7), the following method can also be considered instead of directly solving the simultaneous first-order equations of Eq. (3.8.7). For each $i \in I_A(\mathbf{x}_k)$ and with respect to \mathbf{g}_i , seek $\mathbf{y}_{g0}, \mathbf{y}_{gi_1}, \dots, \mathbf{y}_{gi_{|I_A|}}$ so that the equation

$$\mathbf{y}_{gi} = -\mathbf{H}_{\mathcal{L}}^{-1} \mathbf{g}_i \quad (3.8.9)$$

holds. Furthermore, seek λ_{k+1} using Eq. (3.7.12). In this case, if $I_I(\mathbf{x}_k)$ is non-empty, replace $I_A(\mathbf{x}_k) \setminus I_I(\mathbf{x}_k)$ by $I_A(\mathbf{x}_k)$ and solve Eq. (3.7.8) again. From these results, \mathbf{y}_g is obtained through Eq. (3.7.11).

The difference between this method and the gradient method for constrained problems is that $c_A A$ is replaced by $\mathbf{H}_{\mathcal{L}}$. However, the search vector \mathbf{y}_g obtained with this method can be expected to have the characteristics of the Newton method mentioned in Remark 3.5.3 because it uses the Hesse matrices of the cost functions. However, the following issues must be noted.

Remark 3.8.2 (Newton Method for Constrained Problems) The cost function q of Problem 3.8.1 has the Hesse matrices of the constraint functions multiplied by the Lagrange multipliers of the previous step added on. As a result, the Hesse matrix is not used in the constraint conditions. From this, with respect to problems in which Lagrange multipliers satisfying the KKT conditions change a lot and for which the non-linearity of constraint functions is strong, there are cases when no convergence occurs unless the step size is made small enough. \square

If the second-order derivative of a cost function is already obtained as the Hesse gradient, then the Newton method can now be illustrated as follows. In this case, Problem 3.8.1 is replaced with the following problem.

Problem 3.8.3 (Newton Method Using Hesse Gradient) Let $X = \mathbb{R}^d$, $A \in \mathbb{R}^{d \times d}$ and c_a be a positive definite real symmetric matrix and a positive constant. Moreover, let $\mathbf{x}_k \in X$ be a trial point in Problem 3.1.1, and $\lambda_k \in \mathbb{R}^{|I_A|}$ be the Lagrange multiplier satisfying Eq. (3.7.5) to Eq. (3.7.7) (where $k + 1$ replaces k). Furthermore, let $f_0(\mathbf{x}_k)$, $f_{i_1}(\mathbf{x}_k) = 0, \dots, f_{i_{|I_A|}}(\mathbf{x}_k) = 0$ as well as $\mathbf{g}_0(\mathbf{x}_k)$, $\mathbf{g}_{i_1}(\mathbf{x}_k), \dots, \mathbf{g}_{i_{|I_A|}}(\mathbf{x}_k)$, a search vector $\bar{\mathbf{y}}_g$ and the Hesse gradients $\mathbf{g}_{H0}(\mathbf{x}_k, \bar{\mathbf{y}}_g)$, $\mathbf{g}_{Hi_1}(\mathbf{x}_k, \bar{\mathbf{y}}_g), \dots, \mathbf{g}_{Hi_{|I_A|}}(\mathbf{x}_k, \bar{\mathbf{y}}_g)$ be known, and define

$$\mathbf{g}_{H\mathcal{L}}(\mathbf{x}_k, \bar{\mathbf{y}}_g) = \mathbf{g}_{H0}(\mathbf{x}_k, \bar{\mathbf{y}}_g) + \sum_{i \in I_A(\mathbf{x}_k)} \lambda_{ik} \mathbf{g}_{Hi}(\mathbf{x}_k, \bar{\mathbf{y}}_g). \quad (3.8.10)$$

Under these assumptions, find $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{y}_g$ which satisfies

$$q(\mathbf{y}_g) = \min_{\mathbf{y} \in X} \left\{ q(\mathbf{y}) = \frac{1}{2} \mathbf{y} \cdot (c_a A \mathbf{y}) + (\mathbf{g}_0(\mathbf{x}_k) + \mathbf{g}_{H\mathcal{L}}(\mathbf{x}_k, \bar{\mathbf{y}}_g)) \cdot \mathbf{y} + f_0(\mathbf{x}_k) \mid \begin{array}{l} f_i(\mathbf{x}_k) + \mathbf{g}_i(\mathbf{x}_k) \cdot \mathbf{y} \leq 0 \text{ for } i \in I_A(\mathbf{x}_k) \end{array} \right\}$$

for all $\mathbf{y} \in X$. □

In solving Problem 3.8.3, particularly in the part where we employ the method using the search vectors obtained with respect to each cost function, the same algorithm with the Newton method can be applied by using

$$\mathbf{y}_{gi} = -(c_a A)^{-1} (\mathbf{g}_i + \mathbf{g}_{Hi}) \quad (3.8.11)$$

instead of Eq. (3.8.9).

3.8.1 Simple Algorithm

Bearing in mind the ideas above, let us examine a simple algorithm based on the concept of the Newton method with respect to constrained problems. Figure 3.18 shows the flow diagram of the algorithm. Here, a method for seeking \mathbf{y}_{gi} via Eq. (3.8.9) using \mathbf{g}_i for every $i \in I_A(\mathbf{x}_k)$ is used.

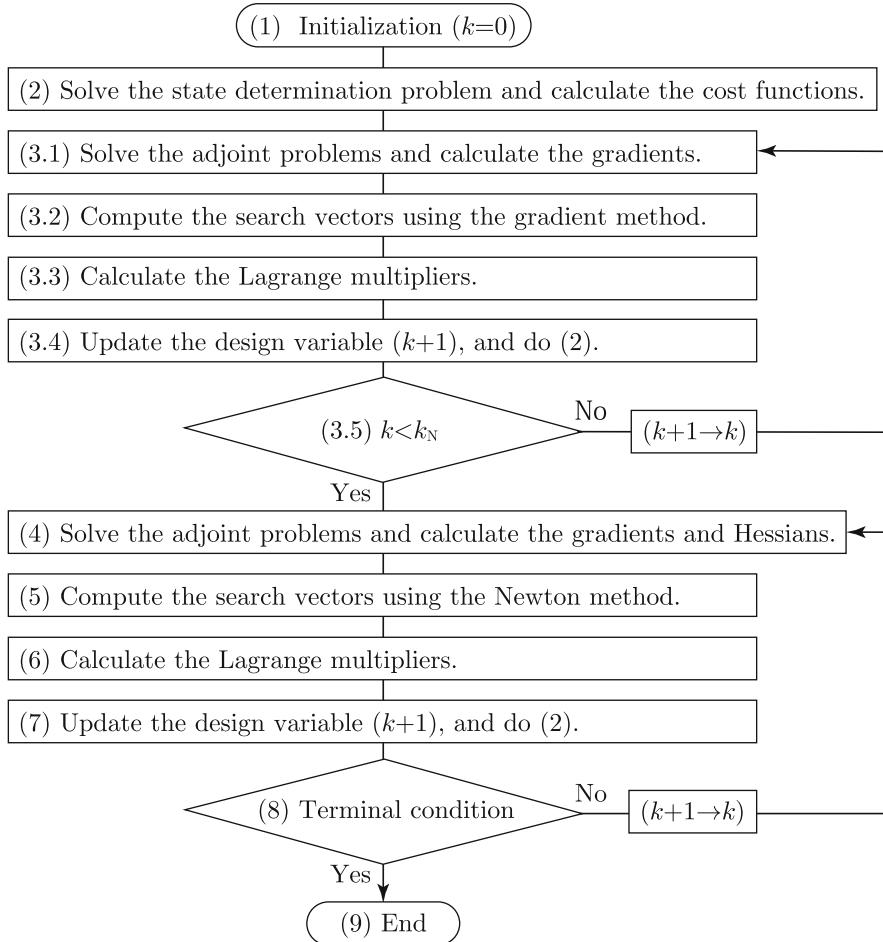


Fig. 3.18 Newton method algorithm with respect to constrained problems

Algorithm 3.8 (Newton Method for Constrained Problems) Obtain the local minimum point of Problem 3.1.1 in the following way:

- (1) Determine the initial point \mathbf{x}_0 so that $f_1(\mathbf{x}_0) \leq 0, \dots, f_m(\mathbf{x}_0) \leq 0$ are satisfied. Determine the positive constant ϵ_0 used for the convergence check of f_0 and the positive constants $\epsilon_1, \dots, \epsilon_m$ which give the permissible range of f_1, \dots, f_m . Set an iteration number k_N at which the Newton method starts, and $k = 0$.
- (2) Solve the state determination problem and calculate $f_0(\mathbf{x}_k), f_1(\mathbf{x}_k), \dots, f_m(\mathbf{x}_k)$. Moreover, define

$$I_A(\mathbf{x}_k) = \{i \in \{1, \dots, m\} \mid f_i(\mathbf{x}_k) \geq -\epsilon_i\}.$$

(3) Do the following when $k < k_N$:

- (3.1) Solve the adjoint problem with respect to $f_0, f_{i_1}, \dots, f_{i_{|I_A|}}$ and solve $\mathbf{g}_0, \mathbf{g}_{i_1}, \dots, \mathbf{g}_{i_{|I_A|}}$ at \mathbf{x}_k .
- (3.2) Use Eq. (3.7.10) to calculate $\mathbf{y}_{g0}, \mathbf{y}_{gi_1}, \dots, \mathbf{y}_{gi_{|I_A|}}$.
- (3.3) Use Eq. (3.7.12) to seek λ_{k+1} . If, in Eq. (3.7.9), $I_I(\mathbf{x}_k)$ is non-empty, replace $I_A(\mathbf{x}_k) \setminus I_I(\mathbf{x}_k)$ with $I_A(\mathbf{x}_k)$ and solve Eq. (3.7.12) again.
- (3.4) Use Eq. (3.7.11) to seek \mathbf{y}_g and let $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{y}_g$ in order to calculate $f_0(\mathbf{x}_{k+1}), f_1(\mathbf{x}_{k+1}), \dots, f_m(\mathbf{x}_{k+1})$. Moreover, let

$$I_A(\mathbf{x}_{k+1}) = \{i \in \{1, \dots, m\} \mid f_i(\mathbf{x}_{k+1}) \geq -\epsilon_i\}.$$

(3.5) Substitute $k+1$ into k . When $k < k_N$, revert to (3.1). Otherwise, proceed to (4).

- (4) Solve the adjoint problem with respect to $f_0, f_{i_1}, \dots, f_{i_{|I_A|}}$ and calculate $\mathbf{g}_0, \mathbf{g}_{i_1}, \dots, \mathbf{g}_{i_{|I_A|}}$ and $\mathbf{H}_0, \mathbf{H}_{i_1}, \dots, \mathbf{H}_{i_{|I_A|}}$ at \mathbf{x}_k .
- (5) Calculate $\mathbf{y}_{g0}, \mathbf{y}_{gi_1}, \dots, \mathbf{y}_{gi_{|I_A|}}$ using Eq. (3.8.9).
- (6) Use Eq. (3.7.12) to seek λ_{k+1} . If $I_I(\mathbf{x}_k)$ in Eq. (3.7.9) is non-empty, replace $I_A(\mathbf{x}_k) \setminus I_I(\mathbf{x}_k)$ with $I_A(\mathbf{x}_k)$ and solve Eq. (3.7.12) again.
- (7) Use Eq. (3.7.11) to seek \mathbf{y}_g . Let $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{y}_g$ and solve the state determination problem in order to calculate $f_0(\mathbf{x}_{k+1}), f_1(\mathbf{x}_{k+1}), \dots, f_m(\mathbf{x}_{k+1})$. Moreover, let

$$I_A(\mathbf{x}_{k+1}) = \{i \in \{1, \dots, m\} \mid f_i(\mathbf{x}_{k+1}) \geq -\epsilon_i\}.$$

(8) Check the terminal condition $|f_0(\mathbf{x}_{k+1}) - f_0(\mathbf{x}_k)| \leq \epsilon_0$.

- When the terminal condition is satisfied, proceed to (9).
- Otherwise, substitute $k+1$ into k and revert to (4).

(9) End the calculations. □

Let us use Algorithm 3.8 in order to find the trial points for Exercise 1.1.7 in Chap. 1.

Exercise 3.8.4 (Mean Compliance Minimization Problem) Consider Exercise 1.1.7. Let the initial point be $\mathbf{a}_{(0)} = (1/2, 1/2)^\top$ and use Algorithm 3.8 in order to obtain the trial points when $k \in \{0, 1\}$. Here, k_N should be taken as small as possible, and the numerical values required should be determined appropriately. □

Answer The mean compliance $\tilde{f}_0(\mathbf{a})$ and the constraint function $f_1(\mathbf{a})$ with respect to volume are given respectively by Eqs. (3.7.13) and (3.7.14). Moreover, their cross-sectional derivatives are given by Eqs. (3.7.15) and (3.7.16). The Hesse

matrix of $\tilde{f}_0(\mathbf{a})$ is

$$\mathbf{H}_0 = \begin{pmatrix} 8/a_1^3 & 0 \\ 0 & 2/a_2^3 \end{pmatrix}. \quad (3.8.12)$$

Moreover, $\mathbf{H}_1 = \mathbf{0}_{\mathbb{R}^{2 \times 2}}$ and λ_1 is not required in Eq. (3.8.2), so we can take $k_N = 0$. Let us seek numerical values using Algorithm 3.8. The design variable is again denoted by $\mathbf{a}_{(k)}$ for each step number k . The same is the case with $\mathbf{b}_{g0(k)}$, $\mathbf{b}_{g1(k)}$ and $\lambda_{1(k)}$.

- (1) At the initial point $\mathbf{a}_{(0)} = (1/2, 1/2)^\top$, $f_1(\mathbf{a}_{(0)}) = 0$ is satisfied. Let $\epsilon_0 = 10^{-3} \tilde{f}_0(\mathbf{a}_{(0)})$, $\epsilon_1 = 10^{-6}$. Set $k = 0$.
- (2) Equations (3.7.13) and (3.7.14) give $\tilde{f}_0(\mathbf{a}_{(0)}) = 10$ and $f_1(\mathbf{a}_{(0)}) = 0$, respectively. Let $I_A(\mathbf{a}_{(0)}) = \{1\}$.
- (3) Since $k = k_N$, proceed to the next step.
- (4) Equations (3.7.15) and (3.7.16) give $\mathbf{g}_{0(0)} = -(16, 4)^\top$, $\mathbf{g}_{1(0)} = (1, 1)^\top$.
- (5) From Eq. (3.8.9), we get $\mathbf{b}_{g0(0)} = (1/4, 1/4)^\top$, $\mathbf{b}_{g1(0)} = -(1/64, 1/16)^\top$.
- (6) Equation (3.7.12) gives $\lambda_{1(1)} = 6.4$.
- (7) Equation (3.7.11) gives $\mathbf{b}_{g(0)} = (0.15, -0.15)^\top$ and letting $\mathbf{a}_{(1)} = \mathbf{a}_{(0)} + \mathbf{b}_{g(0)} = (0.65, 0.35)^\top$ gives $\tilde{f}_0(\mathbf{a}_{(1)}) = 9.01099$, $f_1(\mathbf{a}_{(1)}) = 0$. Moreover, let $I_A(\mathbf{a}_{(1)}) = \{1\}$.
- (8) Since $|\tilde{f}_0(\mathbf{a}_{(1)}) - \tilde{f}_0(\mathbf{a}_{(0)})| = 0.989011 \geq \epsilon_0 = 0.01$, the terminal condition is not satisfied. Then, set $k = 1$ and revert to (4).
- (4) Equations (3.7.15) and (3.7.16) give $\mathbf{g}_{0(1)} = -(9.46746, 8.16327)^\top$, $\mathbf{g}_{1(1)} = (1, 1)^\top$.
- (5) From Eq. (3.8.9), $\mathbf{b}_{g0(1)} = (0.325, 0.175)^\top$, $\mathbf{b}_{g1(1)} = -(0.0343281, 0.0214375)^\top$ can be obtained.
- (6) Equation (3.7.12) is used to obtain $\lambda_{1(2)} = 8.9661$.
- (7) Equation (3.7.11) gives $\mathbf{b}_{g(1)} = (0.0172107, -0.0172107)^\top$, let $\mathbf{a}_{(2)} = \mathbf{a}_{(1)} + \mathbf{b}_{g(1)} = (0.667211, 0.332789)^\top$ which gives $\tilde{f}_0(\mathbf{a}_{(2)}) = 9.00001$, $f_1(\mathbf{a}_{(2)}) = 1.11022 \times 10^{-16}$. Moreover, let $I_A(\mathbf{a}_{(1)}) = \{1\}$.
- (8) $|\tilde{f}_0(\mathbf{a}_{(2)}) - \tilde{f}_0(\mathbf{a}_{(1)})| = 0.010977 \geq \epsilon_0 = 0.01$ shows that the terminal condition is not satisfied, then substitute 2 into k and revert to (4).

Figure 3.19 shows these results and the succeeding computed values. The calculation terminates at $k = 3$, since $|\tilde{f}_0(\mathbf{a}_{(3)}) - \tilde{f}_0(\mathbf{a}_{(2)})| = 0.0000119968 \leq \epsilon_0 = 0.01$. \square

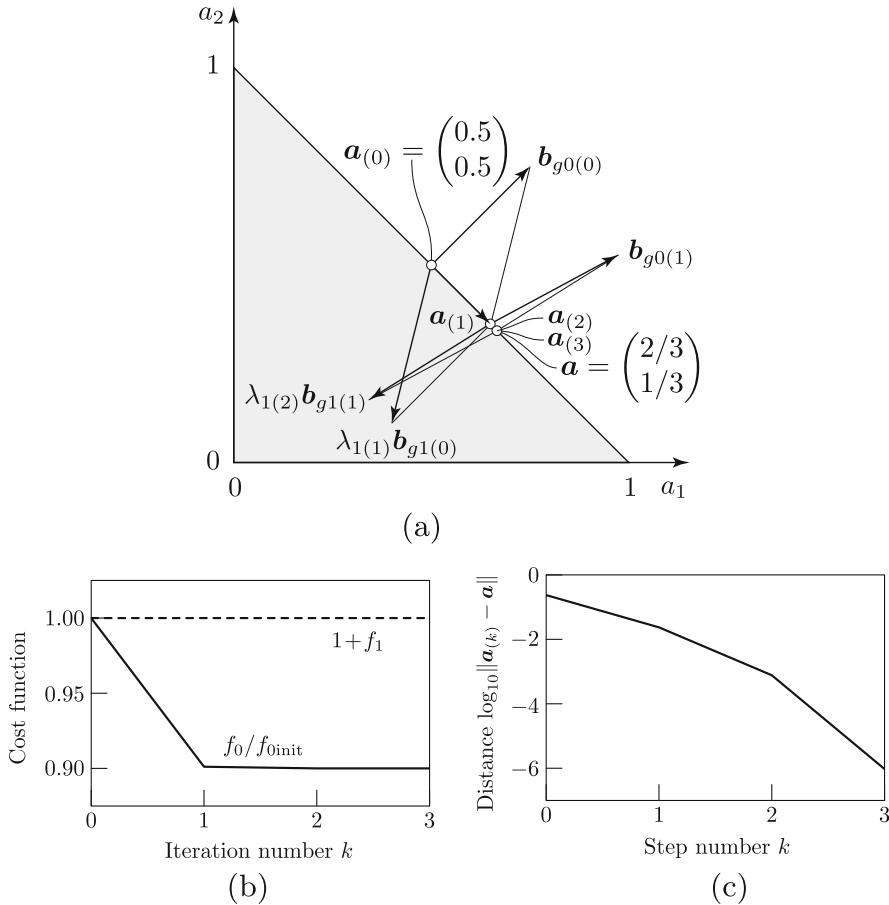


Fig. 3.19 Numerical example of mean compliance minimization problem using Newton method.
 (a) Movement of the trial point. (b) History of cost function. (c) Distance $\|\mathbf{a}_k - \mathbf{a}\|_{\mathbb{R}^2}$

An algorithm via the Newton method using Hesse gradients for the second-order derivatives of cost functions is obtained by replacing Steps (4) and (5) in Algorithm 3.8 as follows:

- (4) Solve the adjoint problems with respect to $f_0, f_{i_1}, \dots, f_{i_{|I_A|}}$ and calculate $\mathbf{g}_0, \mathbf{g}_{i_1}, \dots, \mathbf{g}_{i_{|I_A|}}$. Moreover, solve the adjoint problems with respect to $f'_0, f'_{i_1}, \dots, f'_{i_{|I_A|}}$ and calculate $\mathbf{g}_{H0}, \mathbf{g}_{Hi_1}, \dots, \mathbf{g}_{Hi_{|I_A|}}$.
- (5) Calculate $\mathbf{y}_{g0}, \mathbf{y}_{gi_1}, \dots, \mathbf{y}_{gi_{|I_A|}}$ using Eq. (3.8.11).

Figures 3.20 and 3.21 show the result by the Newton method using the Hesse gradient \mathbf{g}_{H0} of f_0 in Exercise 1.1.7 from the initial point $\mathbf{a}_{(0)} = (1/2, 1/2)^\top$ together with the results by the gradient method (Exercise 3.7.2) and the Newton

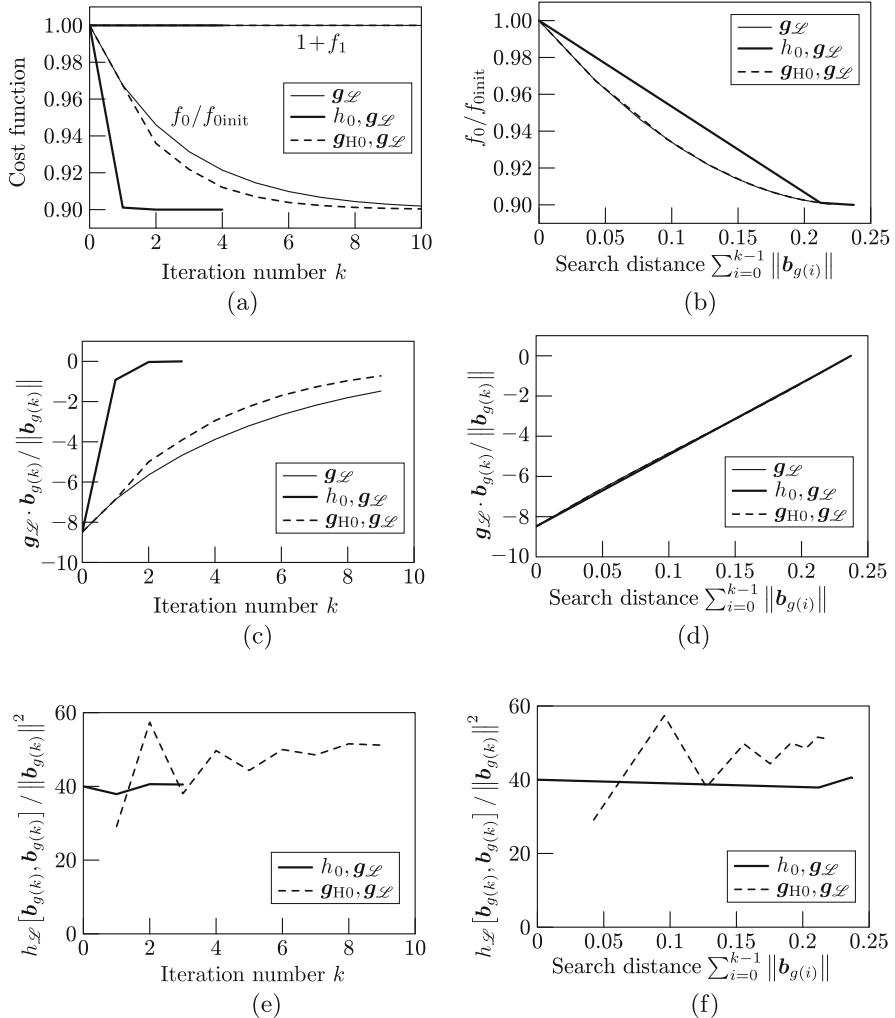


Fig. 3.20 Numerical example of mean compliance minimization: cost functions, gradients and Hessians of f_0 on the search path ($g_{\mathcal{L}}$: Gradient method, $h_0, g_{\mathcal{L}}$: Newton method, $g_{H0}, g_{\mathcal{L}}$: Newton method using Hesse gradient). **(a)** Cost functions. **(b)** Cost functions (search distance). **(c)** Gradient of f_0 on the search path. **(d)** Gradient of f_0 on the search path (search distance). **(e)** Hessian of f_0 on the search path. **(f)** Hessian of f_0 on the search path (search distance)

method (Exercise 3.8.4). In the gradient method, we set $\mathbf{A} = \mathbf{I}$ and the parameter value $c_a = 200$ was assumed. For the Newton method, using the Hesse gradient, we again set $\mathbf{A} = \mathbf{I}$ and chose $c_a = 200$ in the gradient method at $k = 0$ but took $c_a = 100$ for $k \geq k_N = 1$.

Figure 3.20a plots the cost functions $f_0/f_{0\text{init}}$ and $1 + f_1$ normalized with f_0 at the initial shape denoted by $f_{0\text{init}}$ and the volume at the initial shape denoted by $c_1 =$

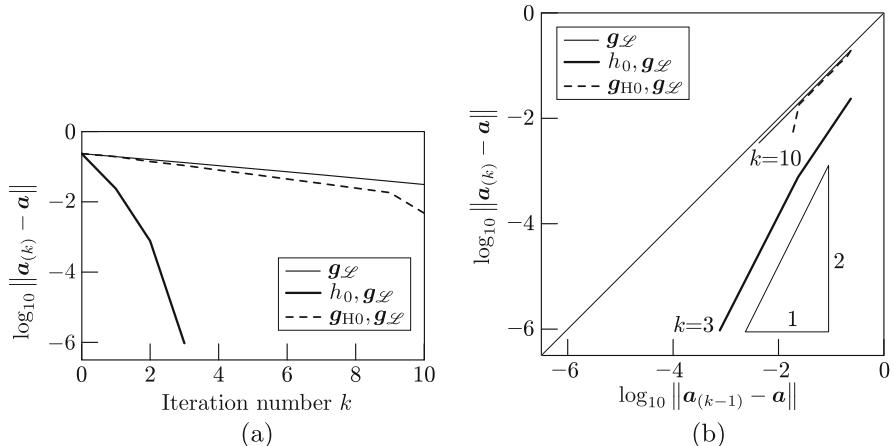


Fig. 3.21 Numerical example of mean compliance minimization: error value $\|\mathbf{a}_k - \mathbf{a}\|_{\mathbb{R}^2}$ between the exact minimum point \mathbf{a} and k -th approximation \mathbf{a}_k ($\mathbf{g}_{\mathcal{L}}$: Gradient method, $\mathbf{h}_0, \mathbf{g}_{\mathcal{L}}$: Newton method, $\mathbf{g}_{\mathbf{H}0}, \mathbf{g}_{\mathcal{L}}$: Newton method using Hesse gradient). (a) Iteration history. (b) $(k-1)$ -th vs. k -th plot

1, respectively, at every iteration number k . Figure 3.20b shows those values with respect to the distance $\sum_{i=0}^k \|\mathbf{b}_{g(i)}\|_X$ on the search path in $X = \mathbb{R}^2$. The graphs of f_0 's gradient (the gradient of the Lagrange function $\mathcal{L} = \mathcal{L}_0 + \lambda_1 f_1$) calculated by $\mathbf{g}_{\mathcal{L}} \cdot \mathbf{b}_{g(k)} / \|\mathbf{b}_{g(k)}\|_X$ are shown in Fig. 3.20c,d with respect to the iteration number and the search distance, respectively. Moreover, Fig. 3.20e,f shows the graphs of f_0 's second-order derivative $h_{\mathcal{L}} [\mathbf{b}_{g(k)}, \mathbf{b}_{g(k)}] / \|\mathbf{b}_{g(k)}\|_X^2$ (in the case of the Newton method using the Hesse gradient, $(\mathbf{g}_{\mathbf{H}0} \cdot \mathbf{b}_{g(k)} + \lambda_1 h_1 [\mathbf{b}_{g(k)}, \mathbf{b}_{g(k)}]) / \|\mathbf{b}_{g(k)}\|_X^2 = \mathbf{g}_{\mathbf{H}0} \cdot \mathbf{b}_{g(k)} / \|\mathbf{b}_{g(k)}\|_X^2$) with respect to the iteration number and the search distance, respectively.

From Fig. 3.20, it can be confirmed that the graphs with respect to the iteration number vary by the difference of the convergence speed, while the graphs with respect to the search distance are almost indistinguishable. The reason is that the search paths are the same as shown in Figs. 3.15a and 3.19a. Such graphs will be shown in Chaps. 8 and 9, too. In these cases, however, it is quite difficult to obtain accurate plots of the search paths so they will no longer be illustrated graphically. Nevertheless, we want the reader to visualize them on their own.

In addition, Fig. 3.21a shows the graphs of the error-norm $\|\mathbf{a}_k - \mathbf{a}\|_X$ between the minimum point \mathbf{a} and the k -th approximation \mathbf{a}_k obtained by the three methods. From this figure, it can be confirmed that the convergence order of the Newton method is higher than the first-order. Moreover, Fig. 3.21b plots the k -th distance $\|\mathbf{a}_k - \mathbf{a}\|_X$ with respect to the $(k-1)$ -th distance $\|\mathbf{a}_{k-1} - \mathbf{a}\|_X$. The indicated slopes of the graphs actually show the convergence order of each method, respectively. This

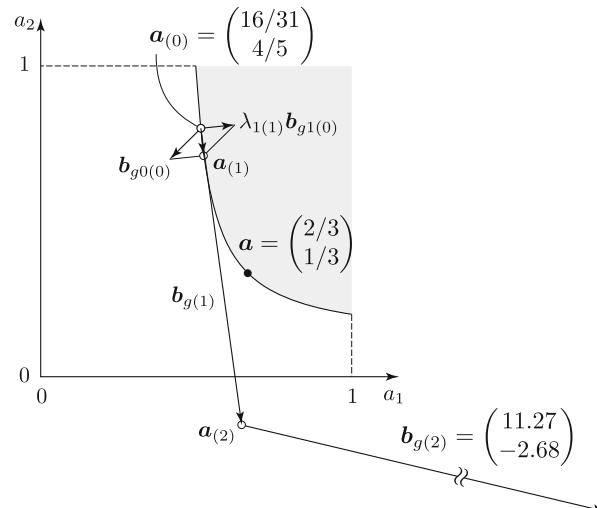
is basically due to the fact that when the equation $\|\mathbf{a}_{(k)} - \mathbf{a}\|_X = r \|\mathbf{a}_{(k-1)} - \mathbf{a}\|_X^p$ is assumed, one also has the relation

$$\log_{10} \|\mathbf{a}_{(k)} - \mathbf{a}\|_X = p \log_{10} \|\mathbf{a}_{(k-1)} - \mathbf{a}\|_X + \log_{10} r. \quad (3.8.13)$$

From the above equation, it is clear that the gradient of the graph (or simply the slope of the graph) corresponds to the order p and the shift of the graph from the diagonal line corresponds to $\log_{10} r$. Based on the slopes of the above plots, we can confirm that the convergence orders of the gradient and Newton method are of first and second order, respectively.

The above sample calculations confirm that the Newton method functions effectively with respect to mean compliance minimization problems. On the other hand, if the Newton method is used with respect to a volume minimization problems, such as Exercise 3.7.3, the resulting values will not reach convergence and will diverge instead. Figure 3.22 shows the movement of a trial point given by Algorithm 3.8. Here, for Lagrange multiplier $\lambda_{1(0)}$ when $k = 0$, $\lambda_{1(1)}$ in Exercise 3.7.3 was used. In this problem, the Hesse matrix of f_0 was $\mathbf{0}_{\mathbb{R}^{2 \times 2}}$ and Hesse matrix of f_1 was positive definite. In other words, the conditions pointed out in Remark 3.8.2 are established. To make it possible to use the Newton method even in situations like this, there is a need to adjust the step size.

Fig. 3.22 Numerical example of a volume minimization problem using the Newton method



3.8.2 *Complicated Algorithm*

If the situation is such as that shown in Fig. 3.22 is considered, then there is a need to add a functionality for adjusting the step size as well as a functionality explained in Sect. 3.7.2. The relationship between these functionalities and algorithm are shown in Sect. 3.7.2, and so, we shall not repeat them here.

Meanwhile, if the Hesse matrix is not positive definite, there are known methods such as making it positive definite by adding a positive definite matrix or making it positive by removing the components of eigenmodes with negative eigenvalues. Furthermore, if it is not positive definite, a gradient method could be used in order to switch to the Newton method once it nears convergence.

3.9 Summary

Chapter 3 looked at methods for seeking the local minimum points with respect to non-linear optimization problems in finite-dimensional vector space. The key points are as follows:

- (1) Iterative methods are used as standard techniques for solving non-linear optimization problems. An iterative method is one in which an initial point is provided and a trial point is updated while appropriately determining a search vector (search direction and step size) (Sect. 3.2).
- (2) A representative method determining the search direction with respect to unconstrained optimization problem is the gradient method. This method is used for determining the search direction defined by the gradient of cost function with respect to the design variable (Sect. 3.3).
- (3) With respect to unconstrained optimization problem, Armijo and Wolfe criteria are known as criteria for determining the appropriateness of the step size. An iterative method, in which the step size has been decided in order to satisfy these criteria, has global convergence (Sect. 3.4).
- (4) If the Hesse matrix and the gradient of the cost function with respect to an unconstrained optimization problem can be calculated, then by using a Newton method, the search direction and step size can be determined simultaneously. The trial point obtained via a Newton method converges quadratically. However, the calculation of the Hesse matrix can be costly (Sect. 3.5).
- (5) The augmented function methods are known as a class of methods for solving optimization problems with inequality constraints. These methods are methods in which the constraint functions are multiplied by a constant representing weight and added to the objective function to make the problem an unconstrained one. However, in order to use these methods, there is a need to find an appropriate monotonic sequence of the constant for each problem (Sect. 3.6).
- (6) A method for solving using KKT conditions can be considered in order to solve optimization problems with inequality constraints. If all the gradients of

cost functions are computable, the gradient method with respect to constrained problems is used. In this method, the Lagrange multipliers are determined using the matrix constructed of search vectors which, on the other hand, are obtained through the gradient method using the gradients for each cost function, as well as the gradients themselves. This relationship is used effectively when considering a practical algorithm (Sect. 3.7).

- (7) If the Hesse matrices of the cost functions in an optimization problem with inequality constraints can be calculated, the Newton method with respect to constrained problems is used. In this method, the positive definite symmetric matrix used in the gradient method with respect to constrained problems is simply replaced by a Hesse matrix and the same algorithm is used as for the gradient method. In order for this method to function effectively, the non-linearity of the constrained functions should be weak (Sect. 3.8).

3.10 Practice Problems

- 3.1** Consider a problem seeking $x \in \mathbb{R}$ which satisfies the gradient $g(x) = 0$ with respect to the non-linear function $f \in C^2(\mathbb{R}; \mathbb{R})$. With respect to $k \in \mathbb{N}$, when $x_k \in \mathbb{R}$ and $g(x_k)$ are given, show the equation for obtaining x_{k+1} with the Newton–Raphson method (Problem 3.5.5). Moreover, show the equation for obtaining x_{k+1} when replacing $g(x_k)$ with the difference

$$\frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}.$$

This demonstrates a formula for the secant method.

- 3.2** Consider using the secant method as an algorithm to solve Problem 3.4.1 (the strict line search method). When $\bar{\epsilon}_{gl}$ is given with respect to $l \in \{0, 1, 2, \dots\}$, show the equation for obtaining $\bar{\epsilon}_{g,l+1}$.

- 3.3** Check that the search direction $\bar{y}_{g,k+1}$ calculated in the equations from (3.4.8) to (3.4.12) shown as an example of a conjugate gradient method (Problem 3.4.9) is conjugate to $\bar{y}_{g,k}$.

- 3.4** Consider a problem seeking the design variable \mathbf{a} (the length of two edges) in Practice 1.6 for which the cost function $f(\mathbf{a})$ (volume of a tetrahedron) is minimized. Obtain the search vector \mathbf{b} using the Newton method when the initial value of the design variable $\mathbf{a}_0 = (a_{01}, a_{02})^\top$ is given.

Chapter 4

Basics of Variational Principles and Functional Analysis



In last chapters, we looked at the theory of solutions relating to optimization problems in finite-dimension. From this chapter onward, we shall consider optimization problems where the design variables are of function of time and space.

We shall first examine variational principles in the field of mechanics and see if they have the structure of a function optimization problem, and if the equation of motion, etc. corresponds to the optimality condition. We can then start to prepare the tools which are needed to deal with function optimization problems. In optimization problems until Chap. 3, the linear space containing the design variables was a finite-dimensional vector space. In contrast, this chapter prepares the function space as the linear space containing design variables. However, in order to be able to explain function spaces, we must start with the definition of a linear space and provide explanations regarding several abstract spaces such as a continuous (complete) distance space in which a limit operation can be used and a linear space in which inner product can be studied. The various function spaces will be explained with consideration to their relationship with the abstract spaces.

Once a function space is defined, we want to consider a mapping between function spaces. Here such a mapping is defined as an operator and its boundedness and linearity will be explained first. The trace operator is an example of an operator and it will be introduced in Chap. 5 and used subsequently when showing the existence of solutions for boundary value problems of partial differential equations and in error evaluation of numerical analysis. After that, operators with a range limited to real numbers are defined and referred to as functionals. Even among such functionals, the set of bounded and linear functionals is referred to as a dual space with respect to the function space which is a domain in the functional. A dual space is a function space used as a gradient in Fréchet derivatives among the generalized derivatives of an operator that will be shown later.

After gathering the tools required in function optimization problems as mentioned above, we return to the variational principles again to completely describe function spaces used in the variational principles. In this regard, variational prin-

ciples are viewed as optimization problems in function spaces. Hence, optimality conditions can be confirmed to be given by the conditions under which the Fréchet differential is zero (KKT conditions in constrained problems). The understanding of variational principles as a function optimization problem will be of immediate use in Chap. 5.

4.1 Variational Principles

Hamilton's principle and minimum principle of potential energy which are well-known in mechanics show that a motion equation or a equilibrium equation of forces can be obtained as a stationary point of some energy. Furthermore, the optimum control law of a control system is known as being obtainable via Pontryagin's minimum principle with respect to an optimum control problem. Here we will look at how these problems have the structure of a function optimization problem.

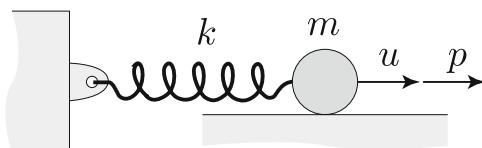
4.1.1 Hamilton's Principle

Let us consider a spring mass system such as one shown in Fig. 4.1. Let k and m be positive constants representing the spring constant and mass. Let t_T be a positive constant representing final time, and $p : (0, t_T) \rightarrow \mathbb{R}$ and $u : (0, t_T) \rightarrow \mathbb{R}$ be functions of time $(0, t_T)$ expressing external force and displacement, respectively. Here, given p , we seek the motion equation for determining u using Hamilton's principle.

With respect to time $t \in (0, t_T)$, let $\dot{u} = \partial u / \partial t$ denote the velocity and functions $\kappa(u, \dot{u})$ and $\pi(u, \dot{u})$ of u and \dot{u} be the kinetic energy and potential energy, respectively. In this case, the Lagrange function in mechanics with respect to the spring mass system of Fig. 4.1 is given by

$$l(u) = \kappa(u, \dot{u}) - \pi(u, \dot{u}) = \frac{1}{2}m\dot{u}^2 - \frac{1}{2}ku^2 + pu. \quad (4.1.1)$$

Fig. 4.1 Spring mass system with one degree of freedom



We have added the phrase “in mechanics” to distinguish it from the Lagrange function used until Chap. 3. Furthermore, the action integral is given by

$$a(u) = \int_0^{t_T} l(u) dt. \quad (4.1.2)$$

Hamilton’s principle forces u to be determined so that $a(u)$ of Eq. (4.1.2) becomes stationary while the displacements $u(0)$ and $u(t_T)$ for times $t = 0$ and $t = t_T$ are fixed with given values. Let us think about the meaning of Hamilton’s principle within the solution of the following problem. In this next problem, however, the terminal condition of Hamilton’s principle has been changed for future use.

Problem 4.1.1 (Extended Hamilton’s Principle) Let α and β be fixed constants, U be the set of $u : (0, t_T) \rightarrow \mathbb{R}$ satisfying $u(0) = \alpha$ and $l(u)$ be Eq. (4.1.1). Moreover, let the expanded action integral be

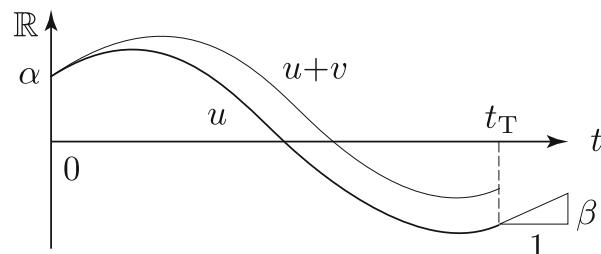
$$f(u) = \int_0^{t_T} l(u) dt - m\beta u(t_T).$$

When u varies arbitrarily within the set U , find the conditions at which f becomes stationary. \square

In Problem 4.1.1, u is an element of the set U of functions satisfying the condition $u(0) = \alpha$. Hence if the function representing an arbitrary variation from u is written as $v : (0, t_T) \rightarrow \mathbb{R}$, there is a need for v to satisfy the condition $v(0) = 0$. Figure 4.2 shows an example of u and v . A set of such v will be expressed as V here. In this case, we have

$$\begin{aligned} f(u+v) &= \int_0^{t_T} \left\{ \frac{1}{2}m(\dot{u} + \dot{v})^2 - \frac{1}{2}k(u+v)^2 + p(u+v) \right\} dt \\ &\quad - m\beta(u(t_T) + v(t_T)) \\ &= f(u) + \left\{ \int_0^{t_T} (m\dot{u}\dot{v} - kuv + pv) dt - m\beta v(t_T) \right\} \end{aligned}$$

Fig. 4.2 Expanded Hamilton’s principle displacement u and arbitrary variation v



$$\begin{aligned}
& + \int_0^{t_T} \left(\frac{1}{2} m \dot{v}^2 - \frac{1}{2} k v^2 \right) dt \\
& = f(u) - \left\{ \int_0^{t_T} (m \ddot{u} + k u - p) v dt - m (\dot{u}(t_T) - \beta) v(t_T) \right\} \\
& \quad + \int_0^{t_T} \left(\frac{1}{2} m \dot{v}^2 - \frac{1}{2} k v^2 \right) dt
\end{aligned} \tag{4.1.3}$$

for every $v \in V$. Here $v(0) = 0$ and partial integration with respect to the integral of $m \dot{u} v$ were used in the last equation. Let us rewrite the right-hand side of this equation collating each term in terms of the order of v :

$$f(u + v) = f(u) + f'(u)[v] + \frac{1}{2} f''(u)[v, v]. \tag{4.1.4}$$

The terms $f'(u)[v]$ and $f''(u)[v, v]$ in Eq. (4.1.4) are referred to as the first variation and second variation of f at u in calculus of variations. The condition for f to be stationary is defined as the condition under which the first variation is zero with respect to an arbitrary $v \in V$. In this problem, this condition is given as

$$f'(u)[v] = - \int_0^{t_T} (m \ddot{u} + k u - p) v dt + m (\dot{u}(t_T) - \beta) v(t_T) = 0. \tag{4.1.5}$$

This condition that holds for an arbitrary $v \in V$ is equivalent to the validity of the following equations:

$$m \ddot{u} + k u = p \quad \text{in } (0, t_T), \tag{4.1.6}$$

$$\dot{u}(t_T) = \beta. \tag{4.1.7}$$

Equations (4.1.6) and (4.1.7) are called the motion equation and terminal condition of velocity, respectively. In Example 4.5.5, it can be verified that $f'(u)[v]$ and $f''(u)[v, v]$ satisfy the definitions of the Fréchet derivative and second-order Fréchet derivative (Definition 4.5.4), respectively. In Sect. 4.6.1, we will look in detail at what sort of function space U is and what type of function space is prepared with respect to p .

4.1.2 Minimum Principle of Potential Energy

Let us consider a one-dimensional linear elastic body such as the one in Fig. 4.3. Let l be a positive constant representing the length, $a_S : (0, l) \rightarrow \mathbb{R}$ and $e_Y : (0, l) \rightarrow \mathbb{R}$ be functions taking positive values representing the cross-section and Young's modulus respectively. Also, let $b : (0, l) \rightarrow \mathbb{R}$, $p_N \in \mathbb{R}$ and $u : (0, l) \rightarrow \mathbb{R}$ denote

the volume force (force per unit volume), the traction (force per unit area) at $x = l$ and displacement, respectively. Here, let us show that when something other than u is given, the minimum principle of potential energy can be used to obtain the equilibrium equation of forces that determines u .

As seen in Example 1.1.1, setting $\pi_I(u)$ and $\pi_E(u)$ as internal potential energy (elastic potential energy) and external potential energy (potential energy of external force) respectively, the potential energy of the whole system when $u = 0$ is the baseline is defined by

$$\begin{aligned}\pi(u) &= \pi_I(u) + \pi_E(u) \\ &= \int_0^l \frac{1}{2} \sigma(u) \varepsilon(u) a_S dx - \int_0^l b u a_S dx - p_N u(l) a_S(l).\end{aligned}\quad (4.1.8)$$

In this case, the strain and stress are set as

$$\begin{aligned}\varepsilon(u) &= \frac{du}{dx} = \nabla u, \\ \sigma(u) &= e_Y \varepsilon(u),\end{aligned}$$

respectively.

Let us obtain the conditions of u from the minimum principle of potential energy with respect to Fig. 4.3.

Problem 4.1.2 (Minimum Principle of Potential Energy) Let U denote the set of all $u : (0, l) \rightarrow \mathbb{R}$ such that $u(0) = 0$ and $\pi(u)$ as Eq. (4.1.8). Find the conditions for u such that

$$\min_{u \in U} \pi(u).$$

□

In Problem 4.1.2, u is an element of the set U of functions satisfying $u(0) = 0$. Denote the function expressing the arbitrary variation from u as $v : (0, l) \rightarrow \mathbb{R}$. In this case, there is a need for v to satisfy $v(0) = 0$. Figure 4.4 shows examples of such u and v . Hence, the set of v is the same as U . Here, we have

$$\begin{aligned}\pi(u + v) &= \int_0^l \frac{1}{2} e_Y (\nabla u + \nabla v)^2 a_S dx - \int_0^l b(u + v) a_S dx \\ &\quad - p_N(u(l) + v(l)) a_S(l)\end{aligned}$$

Fig. 4.3 One-dimensional linear elastic body

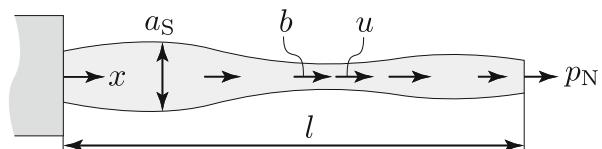
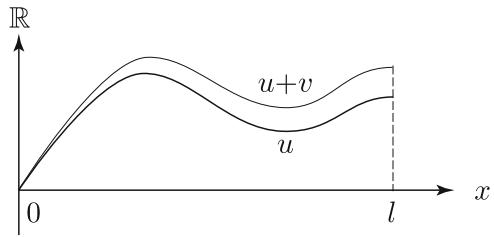


Fig. 4.4 Potential energy minimum principle
displacement u and arbitrary variation v



$$\begin{aligned}
 &= \pi(u) + \left\{ \int_0^l (e_Y \nabla u \nabla v - bv) a_S dx - p_N v(l) a_S(l) \right\} \\
 &\quad + \int_0^l \frac{1}{2} e_Y (\nabla v)^2 a_S dx \\
 &= \pi(u) \\
 &\quad + \left[\int_0^l \{-\nabla(e_Y \nabla u) - b\} v a_S dx + (e_Y \nabla u(l) - p_N) v(l) a_S(l) \right] \\
 &\quad + \int_0^l \frac{1}{2} e_Y (\nabla v)^2 a_S dx
 \end{aligned} \tag{4.1.9}$$

for every $v \in U$, where $v(0) = 0$ and partial integrals relating to the integral of $e_Y \nabla u \nabla v$ were used for the last equation. The right-hand side of this equation is written as

$$\pi(u+v) = \pi(u) + \pi'(u)[v] + \frac{1}{2} \pi''(u)[v, v] \tag{4.1.10}$$

by collating for each order of v . Here the stationary condition of $\pi(u)$ is that the first variation of π ,

$$\pi'(u)[v] = \int_0^l (-\nabla(e_Y \nabla u) - b) v a_S dx + (e_Y \nabla u(l) - p_N) v(l) a_S(l),$$

is zero with respect to an arbitrary $v \in U$. This condition is equivalent to

$$-\nabla(e_Y \nabla u) = -\nabla \sigma(u) = b \quad \text{in } (0, l), \tag{4.1.11}$$

$$\sigma(u(l)) = e_Y \nabla u(l) = p_N. \tag{4.1.12}$$

Furthermore, with respect to the second variation of π , we have

$$\pi''(u)[v, v] = \int_0^l \frac{1}{2} e_Y (\nabla v)^2 a_S dx \geq \alpha \int_0^l (\nabla v)^2 dx \tag{4.1.13}$$

for every $v \in U$. Here $\alpha = \min_{x \in (0, l)} e_Y(x) \min_{x \in (0, l)} a_S(x) / 2 > 0$. From this, the stationary condition of Eqs. (4.1.11) and (4.1.12) express the minimum condition.

Equation (4.1.13) corresponds to the fact that the Hessian matrix of the potential energy was positive definite in Example 2.4.8. In a function optimization problem, the fact that Eq. (4.1.13) is satisfied is expressed as $\pi''(u)[v, v]$ is coercive (Definition 5.2.1).

4.1.3 Pontryagin's Minimum Principle

Finally, let us consider an optimum control problem as an example of using constraints. This section provides suggestions for considering an optimum design problem in the case when time-evolution problems or non-linear problems are set as state determination problems but those in a hurry should skip this section.

Firstly, let us consider the state equations of the system. The motion equation for a system with $n \in \mathbb{N}$ degrees of freedom can be written generally as

$$M\ddot{u} + C\dot{u} + Ku = \xi. \quad (4.1.14)$$

Here M , C and K are seen as matrices in $\mathbb{R}^{n \times n}$ which express the mass, damping and stiffness, respectively. Moreover, $\xi : (0, t_T) \rightarrow \mathbb{R}^n$ and $u : (0, t_T) \rightarrow \mathbb{R}^n$ are seen as the control force and displacement, respectively. Let $v = \dot{u}$. In this case, Eq. (4.1.14) can be rewritten as

$$\begin{pmatrix} I & \mathbf{0}_{\mathbb{R}^{n \times n}} \\ \mathbf{0}_{\mathbb{R}^{n \times n}} & M \end{pmatrix} \begin{pmatrix} \dot{u} \\ v \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{\mathbb{R}^{n \times n}} & -I \\ K & C \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \mathbf{0}_{\mathbb{R}^n} \\ \xi \end{pmatrix}. \quad (4.1.15)$$

In this way the second-order ordinary differential equation with constant coefficient such as Eq. (4.1.14) can be rewritten as a first-order ordinary differential equation with the variable doubled as Eq. (4.1.15). A similar expression also holds for higher-orders.

Let us redefine the symbols, take the control force to be of $d \in \mathbb{N}$ dimensions and define the state determination problem in optimum control problem as follows.

Problem 4.1.3 (Linear System with Control) Let $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times d}$, $\alpha \in \mathbb{R}^n$ and control force $\xi : (0, t_T) \rightarrow \mathbb{R}^d$ be given. Find the system status $u : (0, t_T) \rightarrow \mathbb{R}^n$ which satisfies

$$\dot{u} = Au + B\xi \quad \text{in } (0, t_T), \quad (4.1.16)$$

$$u(0) = \alpha. \quad (4.1.17)$$

□

Having defined the Lagrange function \mathcal{L}_S for the stepped one-dimensional linear elastic problem (Problem 1.1.3) as Eq. (1.1.12), let

$$\mathcal{L}_S(\xi, u, z) = \int_0^{t_T} -(\dot{u} - Au - B\xi) \cdot z \, dt \quad (4.1.18)$$

be the Lagrange function for Problem 4.1.3, where $z : (0, t_T) \rightarrow \mathbb{R}^n$ has been introduced as a Lagrange multiplier with respect to Problem 4.1.3.

Use control force ξ and the solution u of Problem 4.1.3 to set the cost function of optimum control to be

$$f_0(\xi, u) = \frac{1}{2} \int_0^{t_T} \left(\|u\|_{\mathbb{R}^n}^2 + \|\xi\|_{\mathbb{R}^d}^2 \right) dt + \frac{1}{2} \|u(t_T)\|_{\mathbb{R}^n}^2. \quad (4.1.19)$$

Moreover, the constraint of the control force is set as

$$\frac{1}{2} \|\xi\|_{\mathbb{R}^d}^2 - 1 \leq 0 \quad \text{in } (0, t_T). \quad (4.1.20)$$

Here the optimum control problem is constructed as follows.

Problem 4.1.4 (Optimum Control Problem of Linear System) Let Ξ be the set of functions $\xi : (0, t_T) \rightarrow \mathbb{R}^d$ and U be the set of functions $u : (0, t_T) \rightarrow \mathbb{R}^n$. Let f_0 be given by Eq. (4.1.19). In this case, obtain the KKT conditions with respect to ξ satisfying

$$\min_{(\xi, u) \in \Xi \times U} \{f_0(\xi, u) \mid \text{Eq. (4.1.20), Problem 4.1.3}\}. \quad \square$$

Problem 4.1.4 is an optimization problem with equality and inequality constraints and has the same structure as Problem 2.8.1. In Problem 2.8.1, Ξ and U were defined as finite-dimensional vector spaces. Here let us extend Ξ and U to an infinite vector space and formally obtain the KKT conditions.

Define the Lagrange function of Problem 4.1.4 as follows. Let $z_0 : (0, t_T) \rightarrow \mathbb{R}^n$ ($z_0(t_T) = u(t_T)$) be the Lagrange multiplier with respect to Problem 4.1.3 for f_0 and the set of them be expressed as Z . Moreover, let $p : (0, t_T) \rightarrow \mathbb{R}$ be the Lagrange multiplier with respect to Eq. (4.1.20) and the set of these be P . In this case, with respect to an arbitrary $(z_0, p) \in Z \times P$, let

$$\mathcal{L}(\xi, u, z_0, p) = \mathcal{L}_0(\xi, u, z_0) + \mathcal{L}_1(\xi, p) \quad (4.1.21)$$

be the Lagrange function of Problem 4.1.4, where

$$\begin{aligned}\mathcal{L}_0(\xi, \mathbf{u}, z_0) &= f_0(\xi, \mathbf{u}) + \mathcal{L}_S(\xi, \mathbf{u}, z_0) \\ &= \int_0^{t_T} \left\{ \frac{\|\mathbf{u}\|_{\mathbb{R}^n}^2}{2} + \frac{\|\xi\|_{\mathbb{R}^d}^2}{2} - (\dot{\mathbf{u}} - A\mathbf{u} - B\xi) \cdot z_0 \right\} dt \\ &\quad + \frac{1}{2} \|\mathbf{u}(t_T)\|_{\mathbb{R}^n}^2, \end{aligned}\quad (4.1.22)$$

$$\mathcal{L}_1(\xi, p) = \int_0^{t_T} \left(\frac{\|\xi\|_{\mathbb{R}^d}^2}{2} - 1 \right) p dt \quad (4.1.23)$$

are the Lagrange functions with respect to $f_0(\xi, \mathbf{u})$ and Eq. (4.1.20), respectively. Let an arbitrary variation of $(\xi, \mathbf{u}, z_0, p) \in \Xi \times U \times Z \times P$ be $\{\eta, \hat{\mathbf{u}}, \hat{z}_0, \hat{p}\} \in \Xi \times V \times W \times P$, where V is the set of $\hat{\mathbf{u}} : (0, t_T) \rightarrow \mathbb{R}^n$ satisfying $\hat{\mathbf{u}}(0) = \mathbf{0}_{\mathbb{R}^n}$ and W is the set of $\hat{z}_0 : (0, t_T) \rightarrow \mathbb{R}^n$ satisfying $\hat{z}_0(t_T) = \mathbf{0}_{\mathbb{R}^n}$. Then, the first variation of \mathcal{L} with respect to an arbitrary $\{\eta, \hat{\mathbf{u}}, \hat{z}_0, \hat{p}\} \in \Xi \times V \times W \times P$ is

$$\begin{aligned}\mathcal{L}'(\xi, \mathbf{u}, z_0, p) [\eta, \hat{\mathbf{u}}, \hat{z}_0, \hat{p}] &= \mathcal{L}_{0\xi}(\xi, \mathbf{u}, z_0) [\eta] + \mathcal{L}_{1\xi}(\xi, p) [\eta] \\ &\quad + \mathcal{L}_{0\mathbf{u}}(\xi, \mathbf{u}, z_0) [\hat{\mathbf{u}}] + \mathcal{L}_{0z_0}(\xi, \mathbf{u}, z_0) [\hat{z}_0] + \mathcal{L}_{1p}(\xi, p) [\hat{p}]. \end{aligned}\quad (4.1.24)$$

The fourth and fifth terms on the right-hand side of Eq. (4.1.24) are

$$\begin{aligned}\mathcal{L}_{0z_0}(\xi, \mathbf{u}, z_0) [\hat{z}_0] &= - \int_0^{t_T} (\dot{\mathbf{u}} - A\mathbf{u} - B\xi) \cdot \hat{z}_0 dt, \\ \mathcal{L}_{1p}(\xi, p) [\hat{p}] &= \int_0^{t_T} \left(\frac{\|\xi\|_{\mathbb{R}^d}^2}{2} - 1 \right) \hat{p} dt, \end{aligned}$$

respectively. When \mathbf{u} is the solution of the state determination problem (Problem 4.1.3) and Eq. (4.1.20) is satisfied, these terms are zero. Moreover, the third term on the right-hand side of Eq. (4.1.24) is

$$\begin{aligned}\mathcal{L}_{0\mathbf{u}}(\xi, \mathbf{u}, z_0) [\hat{\mathbf{u}}] &= \int_0^{t_T} \left\{ \mathbf{u} \cdot \hat{\mathbf{u}} - (\dot{\hat{\mathbf{u}}} - A\hat{\mathbf{u}}) \cdot z_0 \right\} dt + \mathbf{u}(t_T) \cdot \hat{\mathbf{u}}(t_T) \\ &= \int_0^{t_T} \left(\mathbf{u} + \dot{z}_0 + A^\top z_0 \right) \cdot \hat{\mathbf{u}} dt + (\mathbf{u}(t_T) - z_0(t_T)) \cdot \hat{\mathbf{u}}(t_T), \end{aligned}$$

where $\hat{\mathbf{u}}(0) = \mathbf{0}_{\mathbb{R}^n}$ was used. This term is zero when z_0 is the solution to the following adjoint problem.

Problem 4.1.5 (Adjoint Problem with Respect to f_0) Let $A \in \mathbb{R}^{n \times n}$ be as per Problem 4.1.3. Find $z_0 : (0, t_T) \rightarrow \mathbb{R}^n$ such that

$$\dot{z}_0 = -A^\top z_0 - u \quad \text{in } (0, t_T), \quad (4.1.25)$$

$$z_0(t_T) = u(t_T). \quad (4.1.26)$$

□

The first and the second term on the right-hand side of Eq. (4.1.24) become

$$\begin{aligned} \mathcal{L}_{0\xi}(\xi, u, z_0)[\eta] &= \int_0^{t_T} (\xi + B^\top z_0) \cdot \eta \, dt = \langle g_0, \eta \rangle, \\ \mathcal{L}_{1\xi}(\xi, p)[\eta] &= \int_0^{t_T} p \xi \cdot \eta \, dt = \langle g_1, \eta \rangle, \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes a dual space (Definition 4.4.5). In this case, it is seen as equivalent to the inner product in a finite-dimensional vector space. In addition, in view of the fact that the KKT conditions of Problem 2.8.2 are given by Eq. (2.8.5) to Eq. (2.8.8), the KKT conditions with respect to the solution ξ to Problem 4.1.4 become

$$g_0 + g_1 = (1 + p) \xi + B^\top z_0 = \mathbf{0}_{\mathbb{R}^d} \quad \text{in } (0, t_T), \quad (4.1.27)$$

$$\frac{1}{2} \|\xi\|_{\mathbb{R}^d}^2 \leq 1 \quad \text{in } (0, t_T), \quad (4.1.28)$$

$$\left(\frac{1}{2} \|\xi\|_{\mathbb{R}^d}^2 - 1 \right) p = 0 \quad \text{in } (0, t_T), \quad (4.1.29)$$

$$p \geq 0 \quad \text{in } (0, t_T). \quad (4.1.30)$$

Let us rewrite the KKT condition obtained here into a different expression. Suppose ξ, u, z_0 satisfies Eq. (4.1.27) to Eq. (4.1.30) and $\zeta \in \mathbb{R}^d$ is taken to be an arbitrary vector satisfying $\|\zeta\|_{\mathbb{R}^d}^2 / 2 \leq 1$. In this case, if $\|\xi\|_{\mathbb{R}^d}^2 / 2 < 1$, $p = 0$ is true, then Eq. (4.1.27) implies that

$$\langle g_0, \zeta - \xi \rangle = (\xi + B^\top z_0) \cdot (\zeta - \xi) = 0 \quad \text{in } (0, t_T).$$

Moreover, if $\|\xi\|_{\mathbb{R}^d}^2 / 2 = 1$, then we get $p > 0$, $\xi \cdot (\zeta - \xi) \leq 0$ and

$$\langle g_0 + g_1, \zeta - \xi \rangle = (\xi + B^\top z_0) \cdot (\zeta - \xi) + p \xi \cdot (\zeta - \xi) = 0 \quad \text{in } (0, t_T).$$

Therefore, in any case, the inequality

$$(\xi + B^\top z_0) \cdot (\zeta - \xi) = g_0 \cdot (\zeta - \xi) \geq 0 \quad \text{in } (0, t_T) \quad (4.1.31)$$

holds. Eq. (4.1.31) shows the fact that ξ is a local minimizer of the Lagrange function \mathcal{L}_0 of the cost function f_0 . This condition is called Pontryagin's local minimum condition.

Furthermore, when the Hamilton function is defined as

$$\mathcal{H}(\xi, u, z) = (Au + B\xi) \cdot z + \frac{1}{2} \left(\|u\|_{\mathbb{R}^n}^2 + \|\xi\|_{\mathbb{R}^d}^2 \right),$$

the adjoint problem (Problem 4.1.5) can be written as

$$\begin{aligned} \dot{z}_0 \cdot \hat{u} &= -\mathcal{H}_u(\xi, u, z_0) [\hat{u}] \quad \text{in } (0, t_T), \\ z_0(t_T) &= u(t_T) \end{aligned}$$

with respect to an arbitrary $\hat{u} \in U$. Hence, Eq. (4.1.31) can be written as

$$\mathcal{H}_\xi(\xi, u, z_0) [\zeta - \xi] \geq 0 \quad \text{in } (0, t_T).$$

This relationship shows that

$$\mathcal{H}(\xi, u, z_0) \leq \mathcal{H}(\zeta, v, w_0) \quad \text{in } (0, t_T)$$

with respect to an arbitrary vector ζ satisfying $\|\zeta\|_{\mathbb{R}^d}^2 / 2 \leq 1$. In addition, we see that v and w_0 are solutions to the state determination problem (Problem 4.1.3) and the adjoint problem (Problem 4.1.5), respectively. In this way, ξ which satisfies the KKT conditions of Problem 4.1.4 means that the Hamilton function is minimized. This minimum condition is called Pontryagin's minimum principle.

A similar result can also be obtained for non-linear systems (see [78, Section 5.4, p. 140], where t_T is assumed to be a variable, while being fixed here). The state determination problem with respect to an optimum control problem of non-linear system is defined as follows.

Problem 4.1.6 (Non-linear System with Control) Let $b : \mathbb{R}^d \times \mathbb{R}^n \rightarrow \mathbb{R}^n$, $\alpha \in \mathbb{R}^n$ and control force $\xi : (0, t_T) \rightarrow \mathbb{R}^d$ be given. Find the system status $u : (0, t_T) \rightarrow \mathbb{R}^n$ which satisfies

$$\dot{u} = \frac{\partial b}{\partial u^\top}(\xi, u) u + \frac{\partial b}{\partial \xi^\top}(\xi, u) \xi \quad \text{in } (0, t_T), \quad (4.1.32)$$

$$u(0) = \alpha. \quad (4.1.33)$$

□

Use the control force ξ and solution u of the state determination problem (Problem 4.1.3) to express the cost function of optimum control as

$$f_0(\xi, u) = \int_0^{t_T} h(\xi, u) dt + j(u(t_T)). \quad (4.1.34)$$

Here, $h : \mathbb{R}^d \times \mathbb{R}^n \rightarrow \mathbb{R}$ and $j : \mathbb{R}^n \rightarrow \mathbb{R}$ are given functions. Moreover, when a convex domain $\Omega \subset \mathbb{R}^d$ is given, the constraint of the control force is

$$\xi \in \Omega \quad \text{in } (0, t_T). \quad (4.1.35)$$

The corresponding optimum control problem is constructed as follows.

Problem 4.1.7 (Optimum Control Problem of Non-linear System) Let Ξ be a set of functions $\xi : (0, t_T) \rightarrow \mathbb{R}^d$ and U be a set of functions $u : (0, t_T) \rightarrow \mathbb{R}^n$. Let f_0 be given by Eq. (4.1.34). In this case, obtain the KKT conditions with respect to the problem seeking ξ satisfying

$$\min_{(\xi, u) \in \Xi \times U} \{f_0(\xi, u) \mid \text{Eq. (4.1.35), Problem 4.1.6}\}. \quad \square$$

Define the Lagrange function in a similar way to Problem 4.1.4. Here the adjoint problem of this problem is given as follows.

Problem 4.1.8 (Adjoint Problem with Respect to f_0) With respect to the solution u to Problem 4.1.6 and f_0 of Eq. (4.1.34), obtain $z_0 : (0, t_T) \rightarrow \mathbb{R}^n$ which satisfies

$$\begin{aligned} \dot{z}_0 &= - \left(\frac{\partial \mathbf{b}}{\partial u^\top}(\xi, u) \right)^\top z_0 - \frac{\partial h}{\partial u}(\xi, u), \\ z_0(t_T) &= \frac{\partial j}{\partial u}(u(t_T)). \end{aligned} \quad \square$$

When u and z_0 are the respective solutions to the state determination problem (Problem 4.1.6) and adjoint problem (Problem 4.1.8), the Pontryagin local minimum condition is given as

$$\left(\frac{\partial h}{\partial \xi}(\xi, u) + \left(\frac{\partial \mathbf{b}}{\partial \xi^\top}(\xi, u) \right)^\top z_0 \right) \cdot (\xi - \xi) \geq 0 \quad \text{in } (0, t_T).$$

for an arbitrary $\xi \in \Omega$.

Problems 4.1.4 and 4.1.7 are good examples of how an adjoint problem is constructed when considering the solution of function optimization problem with respect to time-evolution problems.

4.2 Abstract Spaces

In Sect. 4.1, we saw that variational principles have become function optimization problems with functions of time and location as the design variables. In this section and Sect. 4.3, we will look at linear spaces containing design variables of function optimization problems. In this section, we shall start with the definition of a linear space and define abstract spaces that will be needed in future discussions. Here, abstract spaces refer to a set such that rules for the operation or definition for the proximity between all elements are determined.

In this section, a linear space is defined as an abstract space in which linear operators can be applied, and then a metric space is introduced in which metric (distance) can be used. In a metric space, completeness which secures the possibility of limit operations is defined. After that, we return to linear spaces, and linear spaces in which a norm is defined (norm spaces) or linear spaces in which inner products can be applied (inner product spaces) are defined. Norm spaces with completeness (Banach spaces) and inner product space with completeness (Hilbert spaces) are important abstract spaces that will be used in future discussions.

4.2.1 Linear Space

Let us give a definition of a linear space which is one of the more basic abstract spaces in this book. The so-called vector space is another name for a linear space. A linear space is defined as follows.

Definition 4.2.1 (Linear Spaces) With respect to arbitrary elements x and y from set X , let addition $x + y \in X$ be defined. With respect to an arbitrary element α in the set K representing \mathbb{R} or \mathbb{C} and an arbitrary element x in X , let a scalar product $\alpha x \in X$ be defined. In this case, if:

- (1) commutative law of addition $x + y = y + x$,
- (2) associative law of addition $(x + y) + z = x + (y + z)$,
- (3) there exists a zero element $e \in X$ satisfying $e + x = x$,
- (4) there exists an inverse element $-x \in X$ satisfying $(-x) + x = e$,
- (5) there exists a unit element $1 \in K$ satisfying $1x = x$,
- (6) scalar product associative law $\alpha(\beta x) = (\alpha\beta)x$,
- (7) partition law of scalars $(\alpha + \beta)x = \alpha x + \beta x$,
- (8) partition law of vectors $\alpha(x + y) = \alpha x + \alpha y$

holds for every $x, y, z \in X$ and $\alpha, \beta \in K$, then X is called a linear space on K . An element of X is called a vector or a point. An element of K is called a scalar. Furthermore, X is called a real linear space when $K = \mathbb{R}$. \square

With respect to an element x of linear space X , $\alpha x + \beta y$ of an arbitrary $\alpha, \beta \in K$ is called a linear operation or linear combination. Moreover, the direct product of

linear spaces X and Y , $X \times Y$ is a linear space by addition $(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2)$ and scalar product $\alpha (x_1, y_1) = (\alpha x_1, \alpha y_1)$ with respect to arbitrary $(x_1, y_1), (x_2, y_2) \in X \times Y$ and $\alpha \in K$.

Linear spaces have been defined, so let us now provide an example of one. When d is a natural number, it is easy to see that \mathbb{R}^d satisfies the definition of a real linear space. Here, the zero element e is $\mathbf{0}_{\mathbb{R}^d}$ and the inverse element of $x \in \mathbb{R}^d$ is the minus element $-x$.

Set of All Continuous Functions

Next let us look at the fact that the set of all continuous functions with the range of real numbers is a real linear space. From the fact that the set of all continuous functions is a function space, it is appropriate within the contents of this book to provide an explanation that can be found in Sect. 4.3. However, in order to have a concrete image of a linear space, we will provide here a definition of the set of all continuous functions.

From here on, d is taken to be a natural number and k a non-negative integer. Firstly, we will explain a rule known as the multi-index used when expressing partial differentials of continuous functions. With respect to a function which is k -th order partially differentiable $f : \mathbb{R}^d \rightarrow \mathbb{R}$, $\beta = (\beta_1, \dots, \beta_d)^\top \in \{0, \dots, k\}^d$ which satisfies $\sum_{i \in \{1, \dots, d\}} \beta_i \leq k$ is given, $\nabla^\beta f$ and $|\beta|$ are defined as

$$\nabla^\beta f = \frac{\partial^{\beta_1} \partial^{\beta_2} \cdots \partial^{\beta_d} f}{\partial x_1^{\beta_1} \partial x_2^{\beta_2} \cdots \partial x_d^{\beta_d}}, \quad |\beta| = \sum_{i \in \{1, \dots, d\}} \beta_i \leq k$$

respectively. β in this case is called a multi-index. Moreover, the closure of subset $\{x \in \mathbb{R}^d \mid f(x) \neq 0\}$ of \mathbb{R}^d is called the support of f and written as $\text{supp } f$.

The set of all real number functions which is continuous (Sect. A.1.2) up to the k -th order partial derivative is written as follows. Hereinafter, Ω is taken to be \mathbb{R}^d or a connected open subset of \mathbb{R}^d and called a domain (Sect. A.5). When the domain is bounded, its boundary $\partial\Omega$ is assumed to be a Lipschitz boundary (Sect. A.5) and Ω in that case is the Lipschitz domain. Moreover, $\bar{\Omega}$ ($= \Omega \cup \partial\Omega$) expresses the closure of Ω (Sect. A.1.1).

Definition 4.2.2 (Set of All Continuous Functions) Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain. The set of continuous functions $f : \Omega \rightarrow \mathbb{R}$ is defined as follows with respect to $k \in \mathbb{N} \cup \{0\}$:

- (1) The set of all f is written as $C(\Omega; \mathbb{R})$. The set of all f for which $\nabla^\beta f : \Omega \rightarrow \mathbb{R}$ ($|\beta| \leq k$) is continuous is written as $C^k(\Omega; \mathbb{R})$. Furthermore, when the domain is $\bar{\Omega}$, we write $C^k(\bar{\Omega}; \mathbb{R})$.
- (2) The set of all bounded functions f is denoted by $C_B(\Omega; \mathbb{R})$ and the set of such functions $\nabla^\beta f : \Omega \rightarrow \mathbb{R}$ ($|\beta| \leq k$) by $C_B^k(\Omega; \mathbb{R})$.

- (3) The set of all f such that $\text{supp } f$ is a compact set in Ω (see Proposition 4.2.12) is written as $C_0(\Omega; \mathbb{R})$. The set $C^k(\Omega; \mathbb{R}) \cap C_0(\Omega; \mathbb{R})$ is expressed as $C_0^k(\Omega; \mathbb{R})$.

□

The difference between the sets $C(\Omega; \mathbb{R})$ and $C_B(\Omega; \mathbb{R})$ of functions defined in (1) and (2) of Definition 4.2.2 is the boundedness. The set Ω is an open set, hence $C(\Omega; \mathbb{R})$ includes continuous functions which become infinite at the boundary. For example, with respect to $x \in (0, \infty]$, $f(x) = 1/x$ is continuous but is not bounded. In contrast, elements of $C_B(\Omega; \mathbb{R})$ will not include continuous functions which become infinite at the boundary. Thus, we see that

$$C_B(\Omega; \mathbb{R}) \subset C(\Omega; \mathbb{R}). \quad (4.2.1)$$

Moreover, the difference between $C(\bar{\Omega}; \mathbb{R})$ and $C_B(\Omega; \mathbb{R})$ comes from the difference between the defined domain being a closed set $\bar{\Omega}$ or an open set Ω . If the defined domain is a bounded closed set, continuous functions can be shown to be uniformly continuous (Sect. A.1.2) and bounded. On the other hand, if the defined domain is an open set there are examples of them being bounded but not uniformly continuous. For example, for $x \in (0, \infty]$, the function $f(x) = \sin(1/x)$ is continuous and bounded but is not uniformly continuous. Here, the inclusion

$$C(\bar{\Omega}; \mathbb{R}) \subset C_B(\Omega; \mathbb{R}) \quad (4.2.2)$$

is established. However, in a later discussion, a result stating that the norms of $C(\bar{\Omega}; \mathbb{R})$ and $C_B(\Omega; \mathbb{R})$ are the same will be shown (Proposition 4.2.15).

Furthermore, the set $C_0(\Omega; \mathbb{R})$ defined in Definition 4.2.2 (3) shows that it is a set of functions such that $f = 0$ in the neighborhood of the boundary $\partial\Omega$ of Ω (infinity when $\Omega = \mathbb{R}^d$). Here, $C_0^k(\Omega; \mathbb{R})$ is defined in Definition 4.2.2 (3) as a set of functions such that $\nabla^\beta f = 0$ ($|\beta| \leq k$) in the neighborhood of $\partial\Omega$. Based on this definition, we need to emphasize that there is no point in replacing Ω with $\bar{\Omega}$ with respect to $C_0(\Omega; \mathbb{R})$ or $C_0^k(\Omega; \mathbb{R})$. This is because the fact that the support (defined as the closure of domain with non-zero values) of the functions being $\bar{\Omega}$ is inconsistent with the fact that the value of the functions is zero in the neighborhood of the boundary.

Meanwhile, the set $C_0^\infty(\Omega; \mathbb{R})$ defined in Definition 4.2.2 (3) will eventually be used as a test function of Schwartz distribution (Definition 4.3.7) when defining the derivative of a function f within a Sobolev space $W^{k,p}(\Omega; \mathbb{R})$ (Definition 4.3.10). The set $C^\infty(\Omega; \mathbb{R})$ can also be used as a test function when defining the function f which is in Sobolev space $W_0^{k,p}(\Omega; \mathbb{R})$ (Definition 4.3.10) (see Eq. (4.3.10)).

The following can be said about the sets $C^k(\Omega; \mathbb{R})$ and $C_0^k(\Omega; \mathbb{R})$ of all continuous functions.

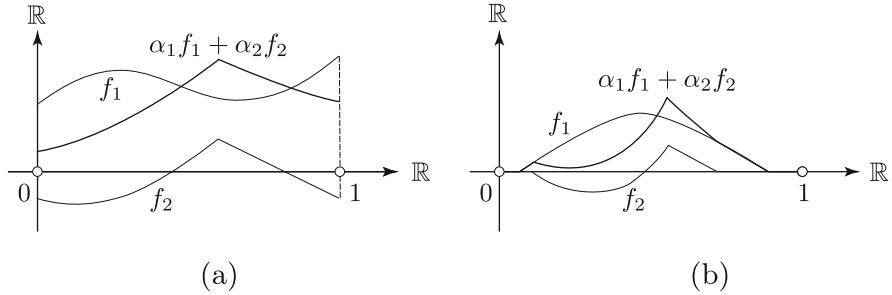


Fig. 4.5 Linear combinations of continuous functions. (a) $f_1, f_2 \in C((0, 1); \mathbb{R})$. (b) $f_1, f_2 \in C_0((0, 1); \mathbb{R})$

Proposition 4.2.3 (Sets of All Continuous Functions) *The sets $C^k(\Omega; \mathbb{R})$, $C^k(\bar{\Omega}; \mathbb{R})$, $C_B^k(\Omega; \mathbb{R})$ and $C_0^k(\Omega; \mathbb{R})$ are real linear spaces with the zero element $f_0 = 0$ in Ω and inverse element $-f$ of f . Moreover, $C^k(\bar{\Omega}; \mathbb{R})$, $C_B^k(\Omega; \mathbb{R})$ and $C_0^k(\Omega; \mathbb{R})$ are real linear subspaces (subsets and real linear spaces) of $C^k(\Omega; \mathbb{R})$.*

□

Proof The fact that a linear combination of continuous functions is a linear function needs to be confirmed. With respect to arbitrary $f_1, f_2 \in C^k(\Omega; \mathbb{R})$ and arbitrary $\alpha_1, \alpha_2 \in \mathbb{R}$, $\alpha_1 f_1 + \alpha_2 f_2$ is an element of $C^k(\Omega; \mathbb{R})$ (see Fig. 4.5a). Therefore, $C^k(\Omega; \mathbb{R})$ is a linear space. The same holds for $C(\bar{\Omega}; \mathbb{R})$ and $C_B^k(\Omega; \mathbb{R})$. Furthermore, $\alpha_1 f_1 + \alpha_2 f_2$ with respect to arbitrary $f_1, f_2 \in C_0^k(\Omega; \mathbb{R})$ and arbitrary $\alpha_1, \alpha_2 \in \mathbb{R}$, $\alpha_1 f_1 + \alpha_2 f_2 = 0$ holds in the neighborhood of $\partial\Omega$, hence $\alpha_1 f_1 + \alpha_2 f_2$ is an element of $C_0^k(\Omega; \mathbb{R})$ (see Fig. 4.5b). Thus, $C_0^k(\Omega; \mathbb{R})$ is a real linear space. In addition, $C^k(\bar{\Omega}; \mathbb{R}) \subset C^k(\Omega; \mathbb{R})$, $C_B^k(\Omega; \mathbb{R}) \subset C^k(\Omega; \mathbb{R})$ and $C_0^k(\Omega; \mathbb{R}) \subset C^k(\Omega; \mathbb{R})$, hence $C^k(\bar{\Omega}; \mathbb{R})$, $C_B^k(\Omega; \mathbb{R})$ and $C_0^k(\Omega; \mathbb{R})$ are real linear subspaces of $C^k(\Omega; \mathbb{R})$. □

Next let us consider the dimension of $C^k(\Omega; \mathbb{R})$. The dimension is defined as follows.

Definition 4.2.4 (Dimension) Let n be a natural number and when a linear space X includes n linear independent vectors but when $n + 1$ vectors are selected they are always linear dependent, then the dimension of X is n . □

Here the following can be said.

Proposition 4.2.5 (Dimension of Set of All Continuous Functions) *The dimension of $C^k(\Omega; \mathbb{R})$ is infinite.* □

Proof We can show that an infinite number of linear independent continuous functions can be found. If $\{f_n\}_{n \in \mathbb{N}} \in C((0, 1); \mathbb{R})$ is selected as $f_1(x) = 1$, $f_2(x) = x$, $f_3(x) = x^2, \dots, f_n(x) = x^{n-1}$, then these are linearly independent.

This is because x^{n-1} cannot be expressed as a linear combination of $1, \dots, x^{n-2}$. Thus, n is infinite since it can be chosen arbitrarily. \square

As seen so far, it has become apparent that the set of all continuous functions is a real linear space (Proposition 4.2.3) of infinite dimension (Proposition 4.2.5).

4.2.2 Linear Subspaces

With the image of a linear space taking shape, we will introduce several definitions relating to the subspaces of linear spaces. First, we will define the linear subspace constructed by linear combination of elements of a finite number as follows.

Definition 4.2.6 (Linear Span) Let m be a natural number and $V = \{x_1, \dots, x_m\}$ be a bounded subset of X with respect to a linear space X in K (\mathbb{R} or \mathbb{C}). Here

$$\text{span } V = \{\alpha_1 x_1 + \dots + \alpha_m x_m \mid \alpha_1, \dots, \alpha_m \in K\}$$

is called a linear span of V , linear hull of V or a linear subspace of X spanned by V . \square

The linear span of V , $\text{span } V$, can be shown to be a subset of X and a minimal linear subspace including V . In the Galerkin method shown as a numerical solution of boundary value problems of partial differential equations in Chap. 6, the set of approximation functions is constructed from a linear span of the given functions.

Moreover, a set constructed as the sum of an element of a linear space and a linear subspace not including the element is referred to as follows.

Definition 4.2.7 (Affine Subspaces) Let X be a linear space and V be a linear subspace of X . With respect to $x_0 \in X \setminus V$, we write

$$V(x_0) = \{x_0 + x \mid x \in V\}$$

and call $V(x_0)$ an affine subspace of V . \square

As an example of an affine subspace, the set U of functions in the expanded Hamiltonian principle (Problem 4.1.1) can be mentioned. As shown later on in Sect. 4.6.1, U is a set of functions u satisfying $u = \alpha$ at $t = 0$ such as in Fig. 4.6a (see Eq. (4.6.3)). In this case U is not a linear space. This is because with respect to $u_1, u_2 \in U$, $u_1 + u_2$ is $2\alpha \neq \alpha$ at $t = 0$. With respect to this, at $t = 0$ such as in Fig. 4.6b, the set of functions V (see Eq. (4.6.4)) satisfying $v = 0$ is a linear space. This is because, in reality, with respect to $v_1, v_2 \in V$, $v_1 + v_2$ is zero at $t = 0$. The condition such that it is zero on a boundary like this is called a homogeneous Dirichlet condition in boundary value problems of partial differential equations. On the other hand, the condition such that it is not zero is called inhomogeneous Dirichlet condition. Here, if $H^1((0, t_T); \mathbb{R})$ (see Definition 4.3.10) which will be

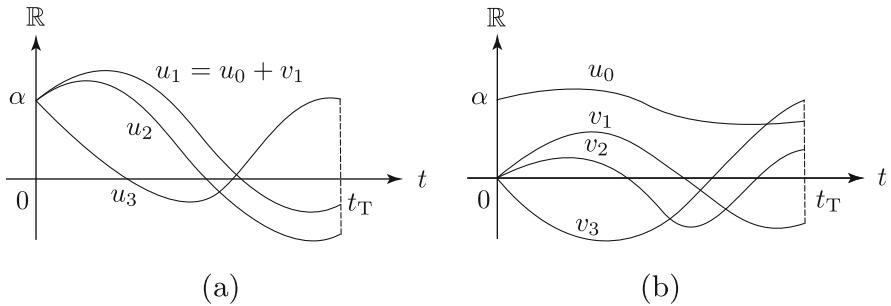


Fig. 4.6 Affine subspace $U = V(u_0)$ of $X = H^1((0, t_T); \mathbb{R})$. (a) $u_1, u_2, u_3 \in U$. (b) $u_0 \in X$, $v_1, v_2, v_3 \in V$

defined later is set to be the linear space X , and a function $u_0 \in X$ (for example u_0 in Fig. 4.6b) which is $u_0 = \alpha$ at $t = 0$ is chosen and fixed, U matches the affine subspace $V(u_0)$ of V . Moreover, $u \in V(u_0)$ agrees with

$$u - u_0 \in V. \quad (4.2.3)$$

In this book, from the fact that importance is given to linear spaces, if a boundary value problem for partial differential equations is to be defined from Chap. 5 onward, the expression Eq. (4.2.3) will mainly be used.

4.2.3 Metric Space

Next, let us think about the conditions which enable limit operation in a set of functions. In this regard, the metric and metric space are defined as follows.

Definition 4.2.8 (Metric Space) With respect to a set X , when a function $d : X \times X \rightarrow \mathbb{R}$ satisfies the following property:

- (1) non-negativity $d(x, y) \geq 0$,
- (2) identity $d(x, y) = 0 \Leftrightarrow x = y$,
- (3) symmetry $d(x, y) = d(y, x)$,
- (4) triangle inequality $d(x, z) \leq d(x, y) + d(y, z)$

for every $x, y, z \in X$, then d is called a metric on X . Moreover, the set X is called metric space with metric d . \square

As can be seen from this definition, a metric space does not need to be a linear space. The characteristics equivalent to the continuity of function spaces should be defined in the metric space. Hence, these definitions will be shown in this section on the metric space.

Densemess

Let us think about the characteristics relating to a subset of a metric space. Let X be a metric space and V be its subset, and \bar{V} be its closure (Sect. A.1.1). In this case, if $X = \bar{V}$, V is referred to as being dense in X . This is equivalent to the fact that there exists at least a point of V in the neighborhood of an arbitrary $x \in X$.

This is explained as followed by using the relationship between the set of all real numbers \mathbb{R} and the set of all rational numbers \mathbb{Q} . When the absolute value $|x - y|$ is taken as its metric with respect to an arbitrary $x, y \in \mathbb{R}$, there exists at least an element of \mathbb{Q} in the neighborhood of an arbitrary $x \in \mathbb{R}$.

Separability

Next, let us think about the quality that an infinite sequence of points can be selected in order to investigate the continuity or closure of metric spaces. Here, infinite sequence of points expresses the set with an infinite number of points from a metric space lined up. When X is a metric space and X has a dense subset constructed of at most a countable number of points (infinite number about the same level as the number of elements in the total set of natural numbers), X is called separable. The set of all rational numbers \mathbb{Q} (because it is a set of fractions with natural numbers as the denominator and the numerator) is a countable set. The set \mathbb{R} contains \mathbb{Q} , hence \mathbb{R} is separable. This type of separability is a characteristic which is a prerequisite when discussing convergence using an infinite sequence of points.

Completeness

The concept of continuity in a metric space is called completeness and is defined using a Cauchy sequence. Firstly, a Cauchy sequence is defined as follows.

Definition 4.2.9 (Cauchy Sequence) If the infinite sequence of points $\{x_n\}_{n \in \mathbb{N}}$ in metric space X satisfies

$$\lim_{n,m \rightarrow \infty} d(x_n, x_m) = 0,$$

then $\{x_n\}_{n \in \mathbb{N}}$ is called a Cauchy sequence. □

The Cauchy sequence is used to define completeness as follows.

Definition 4.2.10 (Completeness) When any Cauchy sequence of a metric space X converges to a point within X , X is called complete. □

In what follows we will explain the meaning of infinite in an infinite sequence of points. It is not conceivable that the meaning of being an infinite number of elements in the set of all natural numbers \mathbb{N} is the same as the meaning of there being an

infinite number of elements in the set of all real numbers \mathbb{R} . It is Cantor's diagonal argument which answered such a simple question (for example, [140, Section 3.4, p. 65]). The concept of cardinality (size) of an infinite set was defined and this showed that the cardinality of \mathbb{R} (cardinality of the continuum) is higher than the cardinality of \mathbb{N} (cardinality of countable set). However, the infinite sequence of points used in the definition of Cauchy sequence is constructed of countable infinity meaning countable cardinality. This can be interpreted that even if the cardinality of \mathbb{R} is cardinality of the continuum, it is sufficient to use a countable infinite number of Cauchy sequences to investigate the completeness in metric spaces.

This can be seen from the fact that the continuity of the set of all real numbers \mathbb{R} is covered by the following axiom.

Axiom 4.2.11 (Completeness of \mathbb{R}) For arbitrary elements $x, y \in \mathbb{R}$, if the absolute value $|x - y|$ is taken as a metric, then every Cauchy sequence in \mathbb{R} converges to a point within \mathbb{R} .

In this axiom every Cauchy sequence in \mathbb{R} can be constructed of just elements of \mathbb{Q} . For example, $\sqrt{2}$ can be defined as a convergent point of an infinite sequence generated by $x_1 = 1$ and $x_{n+1} = x_n/2 + 1/x_n$. The infinite matrix is a Cauchy sequence. In reality, when $n \rightarrow \infty$, $|x_{n+1} - x_n| = |1/x_n - x_n/2| \rightarrow 0$. Therefore, if a set including all the convergence points of Cauchy sequences of \mathbb{Q} is considered, such a set is complete. It is agreed that such a set is seen as \mathbb{R} . In this regard, \mathbb{R} can be said to be a set in which \mathbb{Q} has undergone completion.

The relationship between completeness and subsets of real numbers is as follows. The open interval $(0, 1)$ is not complete. However, the closed interval $[0, 1]$ is complete. This is because the Cauchy sequence $\{1/2, 1/3, 1/4, 1/5, \dots\}$ does not converge within $(0, 1)$, but converges in $[0, 1]$.

Compactness

Denseness indicates the characteristic that even within a subset of a metric space, the closure of the subset is the metric space. In contrast, the characteristic that an infinite sequence from a subset of metric space converging to within that subset is called compactness. Let X be a complete metric space and V its subset. When an arbitrary infinite sequence of points of V includes a partial infinite sequence of points which converges within V , V is said to be compact. When the closure of V contains a convergent infinite subsequence, V is said to be a relative compact set. Here the following proposition is established.

Proposition 4.2.12 (Boundedness of a Compact Set) *Let X be a complete metric space and V be a subset of X . If V is compact, then V is a bounded closed set.* \square

Proof From the definition of compactness, V is a closed set. We will show the boundedness of V by contradiction. If V is not bounded, with respect to a fixed point x in X , there exists an infinite sequence of points $\{y_n\}_{n \in \mathbb{N}}$ that is $d(x, y_n) \rightarrow \infty$. An infinite subsequence of points which converges cannot be selected from within

$\{y_n\}_{n \in \mathbb{N}}$. This is because a convergent infinite sequence of points is bounded. Hence, V needs to be bounded. \square

The reverse proposition of Proposition 4.2.12 “If V is a bounded closed set, then V is compact” is true when X is a finite-dimensional vector space, while it is not true when X is an infinite-dimensional space (see the upper part of Proposition 4.4.11).

4.2.4 Normed Space

Although we have been looking at completeness in metric spaces, there was no need for a metric space to be a linear space. Here we will define a linear space with a defined metric and a linear space with the completeness property.

Let X be a linear space on K . If the function $\|\cdot\| : X \rightarrow \mathbb{R}$ ($\|\mathbf{x}\| : X \ni \mathbf{x} \mapsto \|\mathbf{x}\| \in \mathbb{R}$) has the following properties:

- (1) positivity $\|\mathbf{x}\| \geq 0$,
- (2) equivalence of $\|\mathbf{x}\| = 0$ and $\mathbf{x} = \mathbf{0}$,
- (3) homogeneity or proportionality $\|\alpha \mathbf{x}\| = |\alpha| \|\mathbf{x}\|$,
- (4) triangle inequality $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$

for every $\mathbf{x}, \mathbf{y} \in X$ and arbitrary $\alpha \in K$, then $\|\cdot\|$ is called a norm on X and can also be written as $\|\cdot\|_X$.

In this case, a normed space is defined as follows.

Definition 4.2.13 (Normed Space) A linear space in which a norm is defined is called a normed space. When a set K of scalars is \mathbb{R} , it is called a real normed space. \square

A normed space becomes a metric space if $\|\mathbf{x} - \mathbf{y}\|$ is set as the metric $d(\mathbf{x}, \mathbf{y})$. Therefore, a Cauchy sequence can be defined and completeness can be studied.

Banach Space

A complete normed space is defined as follows.

Definition 4.2.14 (Banach Space) A complete normed space X is called a Banach space. When a set K of scalars is \mathbb{R} , it is called a real Banach space. \square

We will give a specific example. Let the absolute value $|x|$ be a norm with respect to an element x of \mathbb{R} . In this case \mathbb{R} is a Banach space. However, the closed interval $[0, 1]$ is not a Banach space. This is because $[0, 1]$ is not a linear space.

Next, let us consider the set \mathbb{R}^d . For an element $\mathbf{x} = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$, the expression

$$\|\mathbf{x}\|_{\mathbb{R}^d} = \sqrt{|x_1|^2 + \dots + |x_d|^2}$$

is called a Euclidean norm. This norm has the same definition as when norm is defined by $\sqrt{\mathbf{x} \cdot \mathbf{x}}$ using the inner product in \mathbb{R}^d . There are others which satisfy the definition of the norm. For $p \in [1, \infty)$,

$$\|\mathbf{x}\|_p = (|x_1|^p + \cdots + |x_d|^p)^{1/p}$$

is called the p -norm. Meanwhile, for $p = \infty$,

$$\|\mathbf{x}\|_\infty = \max \{|x_1|, \dots, |x_d|\}$$

is called the maximum norm or Chebyshev norm. With respect to these norms, \mathbb{R}^d is a Banach space.

Let us now turn our attention to the set of all continuous functions (Definition 4.2.2). The fact that $C^k(\Omega; \mathbb{R})$ is a real linear space was confirmed by Proposition 4.2.3. However, an element of $C^k(\Omega; \mathbb{R})$ has the possibility of becoming infinite. Hence, when considering a linear space with completeness, $C^k(\Omega; \mathbb{R})$ needs to be excluded. In fact, $C^k(\bar{\Omega}; \mathbb{R})$ and $C_B^k(\Omega; \mathbb{R})$ has the possibility of being a Banach space. Now we will define their corresponding norms and use them to see how completeness (Cauchy sequence of continuous functions converging to continuous functions) can be shown.

In relation to this, it would be good to first investigate how the set of all continuous functions is separable (the fact that Cauchy sequence of continuous functions can be constructed). Continuous functions include the set of all polynomials whose coefficients take rational numbers. There are at most a countable number of these sets. Hence, the set of all polynomials with rational coefficients is a dense subset of the set of all polynomials with real coefficients. Furthermore, from Weierstrass's approximation theorem, the set of all polynomials with real coefficients can be said to be a dense subset of the set of all continuous functions. So it confirms that the set of all continuous functions is separable. Completeness can be verified as follows.

Proposition 4.2.15 (Set of All Continuous Functions) *The sets $C^k(\bar{\Omega}; \mathbb{R})$ and $C_B^k(\Omega; \mathbb{R})$ (Definition 4.2.2) are real Banach spaces with*

$$\|f\|_{C^k(\bar{\Omega}; \mathbb{R})} = \|f\|_{C_B^k(\Omega; \mathbb{R})} = \max_{|\beta| \leq k} \sup_{\mathbf{x} \in \Omega} |\nabla^\beta f(\mathbf{x})|$$

as the norm. □

Proof The key point of the proof is that although the Cauchy sequence of continuous functions converges at each point, it is whether the functions converging at each point are uniformly continuous or not.

Firstly, consider the case when $k = 0$. Suppose $\{f_n\}_{n \in \mathbb{N}} \in C(\bar{\Omega}; \mathbb{R})$ is a Cauchy sequence. From the fact that when fixing on a chosen arbitrary $\mathbf{x} \in \bar{\Omega}$,

$$|f_n(\mathbf{x}) - f_m(\mathbf{x})| \leq \|f_n - f_m\|_{C(\bar{\Omega}; \mathbb{R})} \rightarrow 0$$

with respect to $n, m \rightarrow \infty$ at each point $\mathbf{x} \in \bar{\Omega}$, then there is convergence by \mathbb{R} norm (absolute value). Write this as $f(\mathbf{x})$.

Next, choose n_0 such that, for every $\epsilon > 0$ and $n, m > n_0$, the following holds true:

$$\|f_n - f_m\|_{C(\bar{\Omega}; \mathbb{R})} \leq \frac{\epsilon}{2}.$$

Then, for every $n > n_0$, we have

$$\begin{aligned} |f_n(\mathbf{x}) - f(\mathbf{x})| &\leq |f_n(\mathbf{x}) - f_m(\mathbf{x})| + |f_m(\mathbf{x}) - f(\mathbf{x})| \\ &\leq \|f_n - f_m\|_{C(\bar{\Omega}; \mathbb{R})} + |f_m(\mathbf{x}) - f(\mathbf{x})|. \end{aligned}$$

In this case, if m is taken to be large enough so that $|f_m(\mathbf{x}) - f(\mathbf{x})| \leq \epsilon/2$ is true, $|f_n(\mathbf{x}) - f(\mathbf{x})| \leq \epsilon$ holds and $\{f_n\}_{n \in \mathbb{N}}$ uniformly converges to f . If a continuous function uniformly converges, because of its limit being continuous too, $f \in C(\bar{\Omega}; \mathbb{R})$ and $C(\bar{\Omega}; \mathbb{R})$ is complete.

Now, let $k > 0$. From the definition of the norm of $C^k(\bar{\Omega}; \mathbb{R})$ and with respect to all $|\beta| \leq k$, in the sense of uniform convergence, we have that $\nabla^\beta f_n \rightarrow \nabla^\beta f$ and $\|f_n - f\|_{C^k(\bar{\Omega}; \mathbb{R})} \rightarrow 0$, as $n \rightarrow \infty$, are equivalent. Therefore, $C^k(\bar{\Omega}; \mathbb{R})$ is complete.

Since the elements of $C_B^k(\Omega; \mathbb{R})$ are bounded, the same can be said even when $\bar{\Omega}$ is changed to Ω . \square

The direct product space $X \times Y$ of Banach spaces X and Y becomes a Banach space if the norm $\|(x, y)\|_{X \times Y}$ of $(x, y) \in X \times Y$ is set to be $(\|x\|_X^p + \|y\|_Y^p)^{1/p}$ with respect to $p \in [1, \infty)$ or $\max\{\|x\|_X, \|y\|_Y\}$. Hence, with r as a natural number, the set $C^k(\bar{\Omega}; \mathbb{R}^r)$ of all functions $\mathbf{f} = (f_1, \dots, f_r)^\top : \bar{\Omega} \rightarrow \mathbb{R}^r$ which are k -th order differentiable with \mathbb{R}^r as a range becomes a direct product space $(C^k(\bar{\Omega}; \mathbb{R}))^r$ and if the norm $\|\mathbf{f}\|_{C^k(\bar{\Omega}; \mathbb{R}^r)}$ is $\left(\sum_{i \in \{1, \dots, r\}} \|f_i\|_{C^k(\bar{\Omega}; \mathbb{R})}^p\right)^{1/p}$ or $\max_{i \in \{1, \dots, r\}} \|f_i\|_{C^k(\bar{\Omega}; \mathbb{R})}$, $C^k(\bar{\Omega}; \mathbb{R}^r)$ becomes a Banach space.

Let us confirm here about the need for Banach spaces in optimization problems. If a linear space for which design variables are defined is chosen in a Banach space, the following can be said due to the completeness of Banach spaces. The convergence point when trial points are repeatedly found via iterative methods such as those looked at in Chap. 3 guarantees the existence as an element of a Banach space. Furthermore, in order to use the gradient method, there is a need to generalize the derivative of cost function or gradient method. A generalized derivative is defined for a Banach space in Sect. 4.5 as the Fréchet derivative. A topic of Chap. 7 is the generalization of the gradient method. In this chapter, a Hilbert space which is a complete inner space shown below is required.

4.2.5 Inner Product Space

We will introduce an inner product in a finite-dimensional vector space to an abstract linear space. An inner product can be defined as follows. Let X be a linear space on K (\mathbb{R} or \mathbb{C}). When a function $(\cdot, \cdot) : X \times X \rightarrow K$ satisfies:

- (1) equivalence of $(\mathbf{x}, \mathbf{x}) = 0$ and $\mathbf{x} = \mathbf{0}_X$,
- (2) linearity in a vector $(\mathbf{x} + \mathbf{y}, \mathbf{z}) = (\mathbf{x}, \mathbf{z}) + (\mathbf{y}, \mathbf{z})$,
- (3) linearity in a scalar $(\alpha \mathbf{x}, \mathbf{y}) = \alpha (\mathbf{x}, \mathbf{y})$,
- (4) symmetry $(\mathbf{x}, \mathbf{y}) = (\mathbf{y}, \mathbf{x})$ or Hermitian symmetry $(\mathbf{x}, \mathbf{y}) = (\mathbf{y}, \mathbf{x})^*$ ((\cdot) ^{*} indicates a complex conjugate),
- (5) positivity $(\mathbf{x}, \mathbf{x}) > 0$ for $\mathbf{x} \in X \setminus \{\mathbf{0}\}$

with respect to arbitrary $\mathbf{x}, \mathbf{y}, \mathbf{z} \in X$ and $\alpha \in K$, (\cdot, \cdot) is called an inner product or a scalar product on X . Moreover, (\cdot, \cdot) can be written as $(\cdot, \cdot)_X$. An inner product space is defined as follows.

Definition 4.2.16 (Inner Product Space) A linear space for which an inner product is defined is called an inner product space. When a set K of scalars is \mathbb{R} , it is called a real inner product space. \square

If an inner product is defined, $\|\mathbf{x}\| = \sqrt{(\mathbf{x}, \mathbf{x})}$ satisfies the definition of the norm. Therefore, an inner product space is also a normed space and completeness can be studied.

Hilbert Space

The following can be said about a complete inner product space.

Definition 4.2.17 (Hilbert Space) When an inner product space is complete with respect to the norm

$$\|\mathbf{x}\| = \sqrt{(\mathbf{x}, \mathbf{x})},$$

it is called a Hilbert space. When the set K of scalars is \mathbb{R} , it is called a real Hilbert space. \square

A finite-dimensional vector space \mathbb{R}^d is a real Hilbert space of d dimensions. We will look at the fact that there are Hilbert spaces within function spaces in Sect. 4.3. In fact, the most important abstract space in this book is the Hilbert space. One of the reasons for its importance is that most of the variational principles looked at in Sect. 4.1 have become optimization problems in function spaces where the definition of Hilbert spaces is satisfied. This is confirmed in Sect. 4.6. Moreover, optimum design problems explained in Chap. 7 and beyond will also be defined in a function space in which real Hilbert spaces are defined. Furthermore, the Hilbert space is also used in Chap. 7 when generalizing the gradient method shown in Chap. 3.

4.3 Function Spaces

We have been looking at abstract spaces with completeness and inner products with linear spaces and metric spaces as the base. For these, the set of all continuous functions $C^k(\Omega; \mathbb{R})$ was defined as a set of all continuous functions with value \mathbb{R} defined in Ω . This is to say that it represents the specific set of all functions. A set of all functions whose domain and range satisfy certain conditions such as these is called a function space. Here the function spaces other than the set of all the continuous functions are defined. After all of this we will summarize the definition of norms or inner products if they satisfy the requirements of Banach space or Hilbert space.

Here the domain of the function is written as Ω . However, we mention that this definition can change depending on the function. If a function is continuous, Ω is assumed to be a Lipschitz domain (Sect. A.5) explained prior to Definition 4.2.2. On the other hand, when considering a function for which attention is drawn only to the integral being bounded (integrable), Ω is seen as a measurable set on Ω excluding the subset of \mathbb{R}^d whose Lebesgue measure is zero. However, with respect to $d = 1, 2, 3$, the Lebesgue measure in \mathbb{R}^d will indicate length, area and volume. Hence, the set for which the Lebesgue measure is zero will mean the set of points, length and area respectively with respect to $d = 1, 2, 3$. In equations established on this type of measurable set, a.e. in the sense “almost everywhere” is added.

The proofs relating to theorems and propositions go beyond the level of this book so we refer the interested readers to the literature referenced as an example.

4.3.1 Hölder Space

Firstly, let us define a linear subspace of $C^k(\bar{\Omega}; \mathbb{R})$ where the definition of continuity is even more strict. Lipschitz continuity shown here is used when defining smoothness with respect to domain boundaries (Sect. A.5). In shape optimization problems shown in Chaps. 8 and 9, the fact that design variables vary in order to maintain the smoothness is a topic.

Definition 4.3.1 (Hölder Space) Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain and $f : \bar{\Omega} \rightarrow \mathbb{R}$. If for a $\sigma \in (0, 1]$ there exists some $\beta > 0$ such that

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq \beta \|\mathbf{x} - \mathbf{y}\|_{\mathbb{R}^d}^\sigma, \quad (4.3.1)$$

for every $\mathbf{x}, \mathbf{y} \in \bar{\Omega}$, then f is called Hölder continuous. Here, σ is called the Hölder index and β in this case is called the Lipschitz constant. Moreover, with respect to $k \in \mathbb{N} \cup \{0\}$, the set of f for which $\nabla^\beta f (|\beta| \leq k)$ is Hölder continuous is written

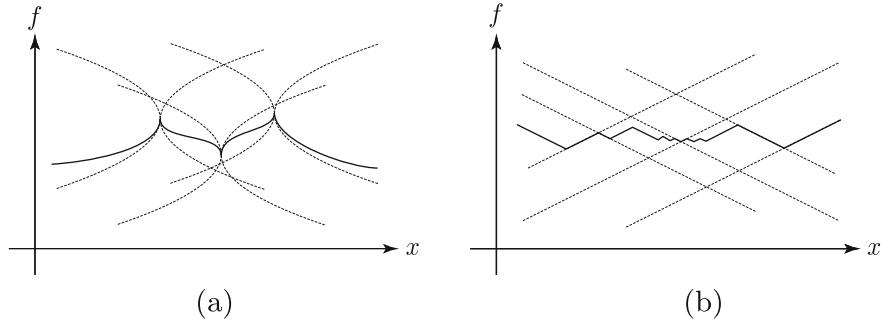


Fig. 4.7 Hölder continuous functions. (a) $\sigma = 0.5$. (b) $\sigma = 1$ (Lipschitz continuous)

as $C^{k,\sigma}(\bar{\Omega}; \mathbb{R})$ and called the Hölder space.¹ In particular, when $k = 0$ and $\sigma = 1$, f is said to be Lipschitz continuous and $C^{0,1}(\bar{\Omega}; \mathbb{R})$ is called the Lipschitz space.

□

Figure 4.7 shows cases when $f : \mathbb{R} \rightarrow \mathbb{R}$ is Hölder continuous and Lipschitz continuous.

The following results can be obtained with respect to $C^{k,\sigma}(\bar{\Omega}; \mathbb{R})$ (cf. [47, Theorem 1, p. 241]).

Proposition 4.3.2 (Hölder Space) *The space $C^{k,\sigma}(\bar{\Omega}; \mathbb{R})$ in Definition 4.3.1 is a real Banach space with*

$$|\nabla^\beta f|_{C^{0,\sigma}(\bar{\Omega}; \mathbb{R})} = \sup_{\mathbf{x}, \mathbf{y} \in \bar{\Omega}, \mathbf{x} \neq \mathbf{y}} \frac{|\nabla^\beta f(\mathbf{x}) - \nabla^\beta f(\mathbf{y})|}{\|\mathbf{x} - \mathbf{y}\|_{\mathbb{R}^d}^\sigma}$$

as the semi-norm and

$$\|f\|_{C^{k,\sigma}(\bar{\Omega}; \mathbb{R})} = \|f\|_{C^k(\bar{\Omega}; \mathbb{R})} + \max_{|\beta|=k} |\nabla^\beta f|_{C^{0,\sigma}(\bar{\Omega}; \mathbb{R})} \quad (4.3.2)$$

as the norm. Here, $\|f\|_{C^k(\bar{\Omega}; \mathbb{R})}$ is defined by Proposition 4.2.15.

□

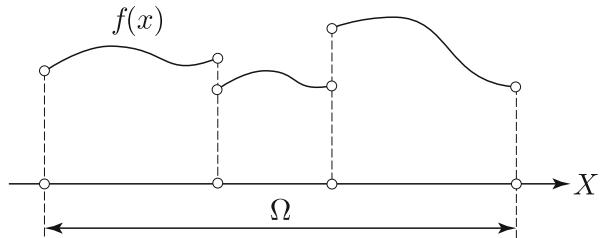
In Eq. (4.3.2), the norm of $C^{k,\sigma}(\bar{\Omega}; \mathbb{R})$ includes the norm of $C^k(\bar{\Omega}; \mathbb{R})$. Therefore, the inclusion

$$C^{k,\sigma}(\bar{\Omega}; \mathbb{R}) \subset C^k(\bar{\Omega}; \mathbb{R}) \quad (4.3.3)$$

holds.

¹According to a book, there are cases when $C^{0,1}(\bar{\Omega}; \mathbb{R})$ is not included in Hölder space.

Fig. 4.8 Integrable functions
 $f : \mathbb{R} \rightarrow \mathbb{R}$



4.3.2 Lebesgue Space

Next we will define the set of all functions for which attention is drawn to the characteristic that integration is defined (integrability) without continuity. Note that integration is defined even for functions such as Fig. 4.8. This sort of integrability is related to energy being defined in variational principles. This can be confirmed in Sect. 4.6.

In Lebesgue spaces and Sobolev spaces defined in this and the next section, the functions $f_1(x)$ and $f_2(x)$ which are the same apart from the set of Lebesgue measure zero are seen as the same functions. This is expressed as $f_1(x) = f_2(x)$ for a.e. $x \in \Omega$. For details see textbooks on Lebesgue integrals.

Definition 4.3.3 (Lebesgue Space) Let $\Omega \subset \mathbb{R}^d$ and $f : \Omega \rightarrow \mathbb{R}$. For $p \in [1, \infty)$, if

$$\int_{\Omega} |f(x)|^p dx < \infty \quad (4.3.4)$$

is satisfied in the sense of a Lebesgue integral (integral of measurable set), f is said to be a p -th order Lebesgue integrable. For $p = \infty$, if

$$\text{ess sup}_{\text{a.e. } x \in \Omega} |f(x)| < \infty \quad (4.3.5)$$

is satisfied, f is said to be essentially bounded. Such a set of all f is called a Lebesgue space and is written as $L^p(\Omega; \mathbb{R})$. \square

The following result can be obtained for $L^p(\Omega; \mathbb{R})$.

Proposition 4.3.4 (Lebesgue Space) The space $L^p(\Omega; \mathbb{R})$ in Definition 4.3.3 is a real Banach space with

$$\|f\|_{L^p(\Omega; \mathbb{R})} = \begin{cases} \left(\int_{\Omega} |f(x)|^p dx \right)^{1/p} & \text{for } p \in [1, \infty), \\ \text{ess sup}_{\text{a.e. } x \in \Omega} |f(x)| & \text{for } p = \infty \end{cases}$$

as the norm. \square

In the proof of completeness of $L^1(\Omega; \mathbb{R})$ of this proposition, Lebesgue's convergence theorem is used (cf. [105, Theorem 2.5, p. 38]). Moreover, the proof that $L^p(\Omega; \mathbb{R})$ when $p \in (1, \infty)$ is a linear space uses Hölder's inequality (Theorem A.9.1) and Minkowski's inequality (Theorem A.9.2) (cf. [105, Theorem 2.10, p. 42]). Furthermore, the linearity and completeness of $L^\infty(\Omega; \mathbb{R})$ can be shown by making the essential supremum $\text{ess sup}_{a.e. x \in \Omega} |f(x)|$ a norm (cf. [105, Theorem 2.20, p. 46]). Here the fact that the norm of $L^\infty(\Omega; \mathbb{R})$ becomes the upper limit value of the absolute value of the function can be understood if it is thought of as an expansion of a maximum value with respect to finite-dimensional vectors to function.

When $p = 2$, $L^2(\Omega; \mathbb{R})$ indicates a set of all functions that are square integrable and is a real Hilbert space with

$$(f, g)_{L^2(\Omega; \mathbb{R})} = \int_{\Omega} f(x) g(x) \, dx \quad (4.3.6)$$

as inner product. This function space is one of the important function spaces in future development.

Here let us look at an example for which a Cauchy sequence of continuous functions can be taken within $L^2(\Omega; \mathbb{R})$ (by $L^2(\Omega; \mathbb{R})$ norm) such that it converges to a discontinuous function which is an element of $L^2(\Omega; \mathbb{R})$.

Exercise 4.3.5 (Cauchy Sequence of Continuous Functions by L^2 Norm) Consider a sequence of functions $\{f_n\}_{n \in \mathbb{N}}$ generated from $C([0, 2]; \mathbb{R})$ by

$$f_n(x) = \begin{cases} x^n & \text{in } (0, 1), \\ 1 - (x - 1)^n & \text{in } (1, 2). \end{cases}$$

Show that $\{f_n\}_{n \in \mathbb{N}}$ is a Cauchy sequence with

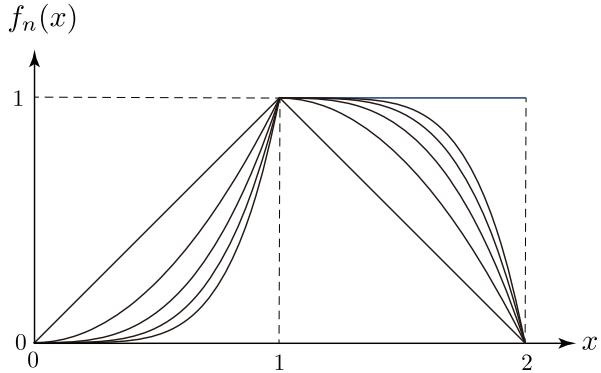
$$\|f\|_{L^2((0, 2); \mathbb{R})} = \left(\int_0^2 |f(x)|^2 \, dx \right)^{1/2}$$

as the norm. Moreover, find the function for which its Cauchy sequence converges. \square

Answer With respect to a function sequence $\{f_n\}_{n \in \mathbb{N}}$ such as the one in Fig. 4.9, when $m, n \rightarrow \infty$, we have

$$\begin{aligned} (f_m, f_n)_{L^2((0, 2); \mathbb{R})} &= \int_0^1 x^{m+n} \, dx + \int_1^2 \{1 - (x - 1)^m\} \{1 - (x - 1)^n\} \, dx \\ &= \frac{1}{1+m+n} + \frac{m}{1+m} - \frac{1}{1+n} - \frac{1}{1+m+n} \rightarrow 1. \end{aligned}$$

Fig. 4.9 Function sequence of continuous functions $\{f_n\}_{n \in \mathbb{N}}$ on $[0, 2]$



Hence, it follows that

$$\begin{aligned} & \|f_m - f_n\|_{L^2((0,2);\mathbb{R})}^2 \\ &= (f_m, f_m)_{L^2((0,2);\mathbb{R})} - 2(f_n, f_m)_{L^2((0,2);\mathbb{R})} + (f_n, f_n)_{L^2((0,2);\mathbb{R})} \rightarrow 0. \end{aligned}$$

Thus, $\{f_n\}_{n \in \mathbb{N}}$ is a Cauchy sequence with respect to the $L^2(\Omega; \mathbb{R})$ norm. Moreover, this Cauchy sequence converges to

$$f = \begin{cases} 0 & \text{in } [0, 1), \\ 1 & \text{in } [1, 2]. \end{cases}$$

In fact, as $n \rightarrow \infty$, we have

$$\begin{aligned} \|f_n - f\|_{L^2((0,1);\mathbb{R})}^2 &= \int_0^1 x^{2n} dx = \frac{1}{1+2n} \rightarrow 0, \\ \|f_n - f\|_{L^2((1,2);\mathbb{R})}^2 &= \int_1^2 \{-(x-1)^n\}^2 dx = \frac{1}{1+2n} \rightarrow 0. \end{aligned}$$

As a result, we get $\|f_n - f\|_{L^2((0,2);\mathbb{R})} \rightarrow 0$ as $n \rightarrow \infty$. □

Based on this fact, it can be interpreted that the completed set by including all functions to which the Cauchy sequence due to $L^2(\Omega; \mathbb{R})$ of $C(\bar{\Omega}; \mathbb{R})$ converges is taken to be $L^2(\Omega; \mathbb{R})$. This is similar to setting the set including all of the convergence points of Cauchy sequence of \mathbb{Q} due to \mathbb{R} norm as \mathbb{R} . Furthermore, it can be also shown that $L^2(\Omega; \mathbb{R})$ is the completion of $C^\infty(\bar{\Omega}; \mathbb{R})$ or $C_0^\infty(\Omega; \mathbb{R})$. These examples are formed by using the Friedrichs mollifier. On the other hand, these facts show that $C(\bar{\Omega}; \mathbb{R})$ becomes a dense subspace of $L^2(\Omega; \mathbb{R})$. From this it can be said that $L^2(\Omega; \mathbb{R})$ is separable (Cauchy sequence can be taken). Separability

can be expanded to $p \in [1, \infty)$ and the separability of $L^p(\Omega; \mathbb{R})$ can be shown by using the fact that $C(\bar{\Omega}; \mathbb{R})$ is a dense subspace of $L^p(\Omega; \mathbb{R})$.

4.3.3 Sobolev Space

In what follows, we will define the set of all integrable functions including derivatives. Variational principles seen in Sect. 4.1 defined energy using integral functions which are displacement differentiated with respect to time or location. Some of those in function space that appear below genuinely have the characteristics required with respect to displacement. This is confirmed in Sect. 4.6. Here we will define the derivative of functions using integrability and then proceed to the main subject.

Schwartz Distribution

Differentiation of functions which are integrable but discontinuous (see Fig. 4.8) utilizes a definition using the Schwartz distribution. Here we will take a look at the definition and the differentiation of discontinuous functions using such a definition.

The definition of the Schwartz distribution uses bounded linear functionals. In this book a bounded linear functional is defined as an operator in Sect. 4.4.5. We will first define it as follows. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a function which is Lebesgue integrable and ϕ be an arbitrary function within $C_0^\infty(\mathbb{R}^d; \mathbb{R})$ (Definition 4.2.2). Here $\langle f, \cdot \rangle : C_0^\infty(\mathbb{R}^d; \mathbb{R}) \rightarrow \mathbb{R}$ defined by

$$\langle f, \phi \rangle = \int_{\mathbb{R}^d} f\phi \, dx < \infty \quad (4.3.7)$$

is called a bounded linear functional determined by f . The arbitrary function $\phi \in C_0^\infty(\mathbb{R}^d; \mathbb{R})$ used here is called a test function since it is a function used experimentally in defining a bounded linear functional. Moreover, the function space of the test function is generally written as $\mathcal{D}(\Omega)$. Hence, hereinafter, $C_0^\infty(\Omega; \mathbb{R})$ will be written as $\mathcal{D}(\Omega)$.

Definition 4.3.6 (Schwartz Distribution) Let $\Omega \subset \mathbb{R}^d$. For the function sequence $\{\phi_n\}_{n \in \mathbb{N}}$ of $\mathcal{D}(\Omega)$ such that when $n \rightarrow \infty$ all the partial derivatives converge uniformly to zero on the compact set in Ω , the bounded linear functional $\langle f, \cdot \rangle : \mathcal{D}(\Omega) \rightarrow \mathbb{R}$ satisfying $\langle f, \phi_n \rangle \rightarrow 0$ ($n \rightarrow \infty$) is called a Schwartz distribution or distribution determined from f . In this book the same symbol f is used for it as long as there is no confusion. The set of Schwartz distributions f with Ω as the domain is written as $\mathcal{D}'(\Omega)$. \square

From Definition 4.3.6, a distribution attempts to define functions which cannot be defined in the ordinary way as a mapping from domain to range in the same way

as bounded linear functionals defined by integrals using test functions with good characteristics.

This type of differentials with respect to the Schwartz distribution is defined below.

Definition 4.3.7 (Partial Derivatives of Schwartz Distributions) Let $\langle f, \cdot \rangle : \mathcal{D}(\Omega) \rightarrow \mathbb{R}$ be a Schwartz distribution determined from f . When

$$\left\langle \frac{\partial f}{\partial x_i}, \phi \right\rangle = - \left\langle f, \frac{\partial \phi}{\partial x_i} \right\rangle, \quad \text{for } i \in \{1, \dots, d\}$$

holds true with respect to an arbitrary $\phi \in \mathcal{D}(\Omega)$, $\langle \partial f / \partial x_i, \cdot \rangle$ is called a partial derivative of a Schwartz distribution or partial derivative of a distribution determined from f . In this book, as long as there is no confusion, the same symbol $\partial f / \partial x_i$ will be used. \square

The fact that $\langle \partial f / \partial x_i, \cdot \rangle$ of Definition 4.3.7 becomes an element of $\mathcal{D}'(\Omega)$ can be said from the fact that even when the function sequence $\{\phi_n\}_{n \in \mathbb{N}}$ of $\mathcal{D}(\Omega)$ is changed to $\{\partial \phi_n / \partial x_i\}_{n \in \mathbb{N}}$ in Definition 4.3.6, $\{\partial \phi_n / \partial x_i\}_{n \in \mathbb{N}}$ is a function sequence of $\mathcal{D}(\Omega)$ (cf. [2, Section 1.60, p. 21], [116, Proposition 2.8, p. 30]). If Definition 4.3.7 is repeatedly used, a higher-order partial derivative in the sense of a Schwartz distribution can be defined. When β is a multi-index and $\nabla^\beta(\cdot) = \partial^{\beta_1} \partial^{\beta_2} \dots \partial^{\beta_d}(\cdot) / \partial x_1^{\beta_1} \partial x_2^{\beta_2} \dots \partial x_d^{\beta_d}$, then

$$\left\langle \nabla^\beta f, \phi \right\rangle = (-1)^{|\beta|} \left\langle f, \nabla^\beta \phi \right\rangle$$

holds. The expression $\langle \nabla^\beta f, \cdot \rangle$ is called a partial derivative of $|\beta|$ -th order in the sense of a Schwartz distribution. In this book, as far as there is no confusion, the same symbol $\nabla^\beta f$ will be used.

Let us mention a specific example here. The function $\delta : \Omega^d \rightarrow \mathbb{R}$ for which

$$\langle \delta, \phi \rangle = \int_{\mathbb{R}^d} \delta \phi \, dx = \phi(\mathbf{0}_{\mathbb{R}^d}) \quad (4.3.8)$$

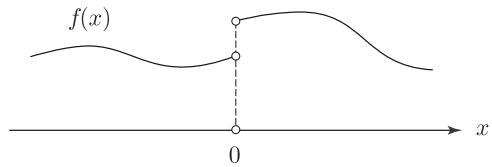
holds with respect to an arbitrary $\phi \in C_0^\infty(\mathbb{R}^d; \mathbb{R})$ is called the Dirac delta function or Dirac distribution. Let us think about the derivative of a step function.

Exercise 4.3.8 (Derivative of Heaviside Step Function) Show that the derivative of the Heaviside step function

$$h = \begin{cases} 0 & \text{in } (-\infty, 0), \\ 1 & \text{in } (0, \infty) \end{cases}$$

as a Schwartz distribution is a Dirac delta function. \square

Fig. 4.10 Discontinuous function $f : \mathbb{R} \rightarrow \mathbb{R}$



Answer From the definition of the derivative of a Schwartz distribution, we have

$$\langle \nabla h, \phi \rangle = \int_{-\infty}^{\infty} \nabla h \phi \, dx = - \int_{-\infty}^{\infty} h \nabla \phi \, dx = - \int_0^{\infty} \nabla \phi \, dx = \phi(0) = \langle \delta, \phi \rangle.$$

□

We have seen how the derivative of the Heaviside step function was a Dirac delta function. Using this relationship, the derivative of a discontinuity function can be written in the following way.

Exercise 4.3.9 (Derivative of Discontinuity Function)

Show the derivative as a Schwartz distribution of a function $f : \mathbb{R} \rightarrow \mathbb{R}$ being discontinuous at the origin, such as the one in Fig. 4.10. □

Answer From the definition of the Schwartz distribution, one obtains

$$\begin{aligned} \langle \nabla f, \phi \rangle &= - \int_{-\infty}^{\infty} f \nabla \phi \, dx = - \int_{-\infty}^0 f \nabla \phi \, dx - \int_0^{\infty} f \nabla \phi \, dx \\ &= (f(0_+) - f(0_-)) \phi(0) + \int_{-\infty}^{\infty} \nabla f \phi \, dx \\ &= (f(0_+) - f(0_-)) \langle \delta, \phi \rangle + \int_{-\infty}^{\infty} \nabla f \phi \, dx, \end{aligned}$$

where $f(0_-) = \lim_{\epsilon \rightarrow 0} f(-\epsilon)$ and $f(0_+) = \lim_{\epsilon \rightarrow 0} f(\epsilon)$ with respect to $\epsilon > 0$. □

Sobolev Space

Since the derivative of an integrable function has been defined, let us now define the function space of integrable functions including derivatives (cf. [56, Definition 1.3.2.1 and Definition 1.3.2.2, p. 16], [116, Definition 9.10, p. 195]).

Definition 4.3.10 (Sobolev Space) Let $\Omega \subset \mathbb{R}^d$. For $k \in \mathbb{N} \cup \{0\}$, $s = k + \sigma$ ($\sigma \in (0, 1)$) and $p \in [1, \infty]$, the entire set of $f : \Omega \rightarrow \mathbb{R}$ such as the following is called a Sobolev space.

- (1) If $\nabla^\beta f \in L^p(\Omega; \mathbb{R})$ holds for all $|\beta| \leq k$, then we write the set of all such functions as $W^{k,p}(\Omega; \mathbb{R})$.
- (2) If $f \in W^{k,p}(\Omega; \mathbb{R})$ holds with respect to $p \in (1, \infty)$ and

$$\int_{\Omega} \int_{\Omega} \frac{|\nabla^\beta f(x) - \nabla^\beta f(y)|^p}{\|x - y\|_{\mathbb{R}^d}^{d+\sigma p}} dx dy < \infty \quad (4.3.9)$$

holds with respect to $|\beta| \leq k$, then we write the set of all such functions as $W^{s,p}(\Omega; \mathbb{R})$.

- (3) The closure of $C_0^\infty(\Omega; \mathbb{R})$ (Definition 4.2.2) in $W^{k,p}(\Omega; \mathbb{R})$ is expressed as $W_0^{k,p}(\Omega; \mathbb{R})$.
- (4) If $k = 0$, then we write the set of all such functions as $W_0^{0,p}(\Omega; \mathbb{R}) = L^p(\Omega; \mathbb{R})$, where $p \in [1, \infty)$ [2, Corollary 2.30, p. 38].
- (5) When $p = 2$, we denote $W^{k,2}(\Omega; \mathbb{R})$ and $W_0^{k,2}(\Omega; \mathbb{R})$ as $H^k(\Omega; \mathbb{R})$ and $H_0^k(\Omega; \mathbb{R})$, respectively. \square

In Definition 4.3.10, with respect to the definition of the Schwartz distribution of $f \in W_0^{k,p}(\Omega; \mathbb{R})$, $C^\infty(\Omega; \mathbb{R})$ is selected as a test function (note that it is not $C_0^\infty(\Omega; \mathbb{R})$). In other words, it is defined via $\langle f, \cdot \rangle : C^\infty(\Omega; \mathbb{R}) \rightarrow \mathbb{R}$ which satisfies

$$\langle f, \phi \rangle = \int_{\Omega} f \phi dx \quad (4.3.10)$$

with respect to an arbitrary $\phi \in C^\infty(\Omega; \mathbb{R})$, instead of Eq. (4.3.7).

The following results can be obtained with respect to $W^{k,p}(\Omega; \mathbb{R})$ (cf. [2, Theorem 3.3, p. 60]).

Proposition 4.3.11 (Sobolev Space) *The space $W^{k,p}(\Omega; \mathbb{R})$ in Definition 4.3.10 is a real Banach space with*

$$\|f\|_{W^{k,p}(\Omega; \mathbb{R})} = \begin{cases} \left(\sum_{|\beta| \leq k} \|\nabla^\beta f\|_{L^p(\Omega; \mathbb{R})}^p \right)^{1/p} & \text{for } p \in [0, \infty), \\ \max_{|\beta| \leq k} \|\nabla^\beta f\|_{L^\infty(\Omega; \mathbb{R})} & \text{for } p = \infty \end{cases} \quad (4.3.11)$$

as the norm. \square

In view of the norm given in Proposition 4.3.11,

$$|f|_{W^{k,p}(\Omega; \mathbb{R})} = \begin{cases} \left(\sum_{|\beta|=k} \left\| \nabla^\beta f \right\|_{L^p(\Omega; \mathbb{R})}^p \right)^{1/p} & \text{for } p \in [0, \infty), \\ \max_{|\beta|=k} \left\| \nabla^\beta f \right\|_{L^\infty(\Omega; \mathbb{R})} & \text{for } p = \infty \end{cases} \quad (4.3.12)$$

is called the semi-norm.

In this book, the space $H^1(\Omega; \mathbb{R})$ is the most important function space. This is because it is a Hilbert space in which the inner product can be used in the following way (cf. [105, Theorem 6.28, p. 134]).

Proposition 4.3.12 (Sobolev Space $H^k(\Omega; \mathbb{R})$) *The space $W^{k,2}(\Omega; \mathbb{R}) = H^k(\Omega; \mathbb{R})$ is a real Hilbert space with*

$$(f, g)_{H^k(\Omega; \mathbb{R})} = \sum_{|\beta| \leq k} \int_{\Omega} \nabla^\beta f \cdot \nabla^\beta g \, dx \quad (4.3.13)$$

as the inner product. \square

Among the spaces $H^k(\Omega; \mathbb{R})$, the sets $H^1(\Omega; \mathbb{R})$ and $H^1(\Omega; \mathbb{R}^d)$ are the most important function spaces used as function spaces that the functions from shape or topology optimization problems or boundary value problems of partial differential equation can be described by. Hence, let us define the inner product for validation. The inner product of $f, g \in H^1(\Omega; \mathbb{R})$ can be defined by

$$(f, g)_{H^1(\Omega; \mathbb{R})} = \int_{\Omega} (fg + \nabla f \cdot \nabla g) \, dx. \quad (4.3.14)$$

Moreover, the inner product of $f, g \in H^1(\Omega; \mathbb{R}^d)$ can be defined by

$$\begin{aligned} (f, g)_{H^1(\Omega; \mathbb{R}^d)} &= \int_{\Omega} \left\{ f \cdot g + (\nabla f^\top) \cdot (\nabla g^\top) \right\} \, dx \\ &= \int_{\Omega} \left(f \cdot g + \sum_{(i,j) \in \{0, \dots, d\}^2} \left(\frac{\partial f_i}{\partial x_j} \right)_{ij} \left(\frac{\partial g_i}{\partial x_j} \right)_{ij} \right) \, dx. \end{aligned} \quad (4.3.15)$$

Let us investigate what is and what is not included in $H^1((0, 1); \mathbb{R})$ using a power function.

Exercise 4.3.13 (Power Function in $H^1((0, 1); \mathbb{R})$) Let $x \in (0, 1)$. Determine the conditions on $\alpha \in \mathbb{R}$ such that the function

$$f = x^\alpha$$

is an element of $H^1((0, 1); \mathbb{R})$. \square

Answer Let the derivative of f with respect to x be written as f' . Here, in order for the following inequality to hold,

$$\begin{aligned} \|f\|_{H^1((0, 1); \mathbb{R})} &= \left\{ \int_0^1 (f^2 + f'^2) dx \right\}^{1/2} = \left\{ \int_0^1 (x^{2\alpha} + \alpha^2 x^{2(\alpha-1)}) dx \right\}^{1/2} \\ &= \left(\left[\frac{x^{2\alpha+1}}{2\alpha+1} + \frac{\alpha^2 x^{2\alpha-1}}{2\alpha-1} \right]_0^1 \right)^{1/2} < \infty, \end{aligned}$$

it must be $2\alpha - 1 > 0$ or equivalently $\alpha > 1/2$. \square

From Exercise 4.3.13, we can see that the singularity (Sect. 5.3) of $f = \sqrt{x}$ at $x = 0$ is not permissible within the space $H^1((0, 1); \mathbb{R})$.

4.3.4 Sobolev Embedding Theorem

According to the definition of Sobolev spaces $W^{k,p}(\Omega; \mathbb{R})$ (Definition 4.3.10), many function spaces can be created depending on how $d \in \mathbb{N}$, $k \in \mathbb{N} \cup \{0\}$ and $p \in [1, \infty]$ are selected. Furthermore, the embedding relationships of various function spaces including the Hölder space $C^{k,\sigma}(\Omega; \mathbb{R})$ are summarized by the Sobolev embedding theorem, shown below.

Let us start by taking a general view on such an embedded relationship. Suppose $\Omega \subset \mathbb{R}^d$ and k is fixed. If $q < p$, then the inclusion $W^{k,p}(\Omega; \mathbb{R}) \subset W^{k,q}(\Omega; \mathbb{R})$ holds. Moreover, if p is fixed, then $W^{k+1,p}(\Omega; \mathbb{R}) \subset W^{k,p}(\Omega; \mathbb{R})$ also holds. These relationships are apparent from the definition of the norm. The Sobolev embedding theorem shows the embedding relationship between Sobolev spaces with different p and k . This shows that when $k - d/p$ is viewed as an order of the differentiability and

$$k + 1 - \frac{d}{p} \geq k - \frac{d}{q}$$

holds, then the inclusion

$$W^{k+1,p}(\Omega; \mathbb{R}) \subset W^{k,q}(\Omega; \mathbb{R})$$

is true. Furthermore, if $0 < \sigma = k - d/p < 1$, it shows that

$$W^{k,p}(\Omega; \mathbb{R}) \subset C^{0,\sigma}(\Omega; \mathbb{R})$$

holds. In this case, there is a need for Ω to be a Lipschitz domain.

With these relationships in mind, let us look at a detailed explanation of Sobolev embedding theorem (cf. [2, Theorem 4.12, p. 85]).

Theorem 4.3.14 (Sobolev Embedding Theorem) *Let $\Omega \subset \mathbb{R}^d$. Let $k \in \mathbb{N}, j \in \mathbb{N} \cup \{0\}$ and $p \in [1, \infty)$. Then, the following results hold:*

(1) *if $k - d/p < 0$, with $p^* = d / \{(d/p) - k\}$, then*

$$W^{k+j,p}(\Omega; \mathbb{R}) \subset W^{j,q}(\Omega; \mathbb{R}) \quad \text{for } q \in [p, p^*], \quad (4.3.16)$$

(2) *if $k - d/p = 0$, then*

$$W^{k+j,p}(\Omega; \mathbb{R}) \subset W^{j,q}(\Omega; \mathbb{R}) \quad \text{for } q \in [p, \infty), \quad (4.3.17)$$

(3) *if $k - d/p = j + \sigma > 0$ ($\sigma \in (0, 1)$), or $k = d$ and $p = 1$, then*

$$W^{k+j,p}(\Omega; \mathbb{R}) \subset W^{j,q}(\Omega; \mathbb{R}) \quad \text{for } q \in [p, \infty]. \quad (4.3.18)$$

Furthermore, if Ω is a Lipschitz domain, then

(4) *if $k - d/p = j + \sigma > 0$ ($\sigma \in (0, 1)$), or $k = d$ and $p = 1$, we have*

$$W^{k+j,p}(\Omega; \mathbb{R}) \subset C^{j,\lambda}(\bar{\Omega}; \mathbb{R}) \quad \text{for } \lambda \in (0, \sigma], \quad (4.3.19)$$

(5) *if $k - 1 = d$ and $p = 1$, then we have*

$$W^{k+j,p}(\Omega; \mathbb{R}) \subset C^{j,1}(\bar{\Omega}; \mathbb{R}). \quad (4.3.20)$$

□

In order to understand the idea on how Theorem 4.3.14 (1) holds, we use a Sobolev inequality. Under the assumption of Theorem 4.3.14 (1), Sobolev inequality theorem is provided by

$$\|f\|_{L^q(\Omega; \mathbb{R})} \leq c |f|_{W^{k,p}(\Omega; \mathbb{R})} \quad (4.3.21)$$

with respect to an arbitrary $f \in C_0^\infty(\Omega; \mathbb{R})$, where c is a positive constant which is not dependent on f . Here, we pay attention to Eq. (4.3.21) when $k = 1$ (see, for example, [2, Theorem 4.31, p. 102], [32, Théorème IX.9, p. 162]).

Let $a > 0$. For an element $\mathbf{x} \in \Omega$, let $\mathbf{y} = a\mathbf{x} \in \hat{\Omega}$ and $f(\mathbf{x}) = \hat{f}(\mathbf{y})$. In this case, with respect to the left-hand side of Eq. (4.3.21), we have

$$\begin{aligned}\|f\|_{L^q(\Omega; \mathbb{R})} &= \left(\int_{\Omega} |f|^q \, dx_1 \cdots dx_d \right)^{1/q} = a^{-d/q} \left(\int_{\hat{\Omega}} |\hat{f}|^q \, dy_1 \cdots dy_d \right)^{1/q} \\ &= a^{-d/q} \|\hat{f}\|_{L^q(\hat{\Omega}; \mathbb{R})}.\end{aligned}\quad (4.3.22)$$

On the other hand, with respect to $|f|_{W^{k,p}(\Omega; \mathbb{R})}$ in the right-hand side of Eq. (4.3.21), we get that

$$\begin{aligned}|f|_{W^{1,p}(\Omega; \mathbb{R})} &= \left(\int_{\Omega} \sum_{|\beta|=1} \left| \frac{\partial^{|\beta|} f}{\partial x_1^{\beta_1} \cdots \partial x_d^{\beta_d}} \right|^p \, dx_1 \cdots dx_d \right)^{1/p} \\ &= a^{(p-d)/p} \left(\int_{\hat{\Omega}} \sum_{|\beta|=1} \left| \frac{\partial^{|\beta|} \hat{f}}{\partial y_1^{\beta_1} \cdots \partial y_d^{\beta_d}} \right|^p \, dy_1 \cdots dy_d \right)^{1/p} \\ &= a^{1-d/p} |\hat{f}|_{W^{1,p}(\hat{\Omega}; \mathbb{R})}\end{aligned}\quad (4.3.23)$$

holds. Here, note that the assumption on Theorem 4.3.14 (1) that $q \leq p^* = d / \{(d/p) - 1\}$ is written as

$$\frac{1}{p^*} = \frac{1}{p} - \frac{1}{d} \leq \frac{1}{q},$$

and it can also be written as $1 - d/p + d/q \geq 0$. Hence, if Eq. (4.3.21) holds, then for an arbitrary $a > 0$, the inequality

$$\|\hat{f}\|_{L^q(\hat{\Omega}; \mathbb{R})} \leq a^{1-d/p+d/q} c |\hat{f}|_{W^{1,p}(\hat{\Omega}; \mathbb{R})} \quad (4.3.24)$$

holds. Since if $1 - d/p + d/q < 0$, then $a^{1-d/p+d/q} \rightarrow 0$ when $a \rightarrow \infty$, the assumption of Theorem 4.3.14 (1) is needed in order that Eq. (4.3.21) with respect to \hat{f} holds.

Moreover, in Theorem 4.3.14 (4) and (5), the embedding relationship of the Sobolev space $W^{k,p}(\Omega; \mathbb{R})$ and Hölder space $C^{0,\sigma}(\bar{\Omega}; \mathbb{R})$ ($\sigma \in (0, 1]$) is given. The relationship between the two needs to be explained. This is because in contrast to functions included in $C^{0,\sigma}(\bar{\Omega}; \mathbb{R})$ which are functions with values on all points of $\bar{\Omega}$, functions included in $W^{k,p}(\Omega; \mathbb{R})$ are functions defined on a measurable set of Ω (almost everywhere). If comparing both under these definitions, it is considered that

$f \in W^{k,p}(\Omega; \mathbb{R})$ and an equivalent function $f^* \in C^{0,\sigma}(\bar{\Omega}; \mathbb{R})$ such that $f = f^*$ holds on the measurable set can be selected and

$$\|f^*\|_{C^{0,\sigma}(\bar{\Omega}; \mathbb{R})} \leq c \|f\|_{W^{k,p}(\Omega; \mathbb{R})} \quad (4.3.25)$$

is established with respect to some $c > 0$ (cf. [2, Section 4.2, p. 79]).

Even among Hölder spaces, with respect to the embedding relationship between the Lipschitz space $C^{0,1}(\bar{\Omega}; \mathbb{R})$ and Sobolev space $W^{1,\infty}(\Omega; \mathbb{R})$, $W^{1,\infty}(\Omega; \mathbb{R}) = C^{0,1}(\bar{\Omega}; \mathbb{R})$ is established for the convex set Ω and real-valued (not \mathbb{R}^n) function f . In fact, the norms of $f \in W^{1,\infty}(\Omega; \mathbb{R})$ and its equivalent function $f^* \in C^{0,1}(\bar{\Omega}; \mathbb{R})$ are the same [97, Proposition 1.39, p. 23].

Furthermore, in Theorem 4.3.14, k and j were assumed to be integers. It is known that the following relationship holds with respect to embedding relationships when these are expanded to real numbers s and t (cf. [56, Eq. (1.4.4.5) and Eq. (1.4.4.6), p. 27]). Suppose s and $t \leq s$ are non-negative real numbers and p and $q \geq p$ are defined by the relationship when k is replaced by s in Theorem 4.3.14. Here, if

$$s - \frac{d}{p} \geq t - \frac{d}{q} \quad (4.3.26)$$

is satisfied, we get

$$W^{s,p}(\Omega; \mathbb{R}) \subset W^{t,q}(\Omega; \mathbb{R}). \quad (4.3.27)$$

Furthermore, when Ω is a Lipschitz domain and $k < k + \sigma = s - d/p < k + 1$ (k is a non-negative integer), we get

$$W^{s,p}(\Omega; \mathbb{R}) \subset C^{k,\sigma}(\Omega; \mathbb{R}). \quad (4.3.28)$$

4.4 Operators

In Sect. 4.3, various function spaces were defined and we looked at how these become a Banach space or a Hilbert space. In function optimization problems, it can be said that we had been looking at linear spaces involving design variables. In optimum design problems with functions as design variables, it can also be linear spaces containing the state variables which are the solution to the state determination problems. Next, we want to examine how the derivative of the cost function is defined when the cost function is given by the integral (functional) consisting of a design variable or a state variable. Here, as a preparation for this, a mapping from between two Banach spaces is defined as an operator. Moreover, operators with range as real numbers are defined as functionals. Furthermore, the set of functionals which domain is a function space is a Banach space and defined as a dual space

with respect to that function space. This dual space is an important function space containing gradients when defining the derivative of a functional in Sect. 4.5. In this section we shall also describe important theorems (trace theorem and Riesz representation theorem) relating to operators other than dual spaces.

4.4.1 Bounded Linear Operator

Performing a calculation on a function is to define a mapping between two function spaces. The mapping in this case in particular is called an operator. An operator with linearity is called a linear operator and can be defined as follows. Let X and Y be Banach spaces on K . With respect to arbitrary $\mathbf{x}_1, \mathbf{x}_2 \in X$ and $\alpha_1, \alpha_2 \in K$, if the mapping $f : X \rightarrow Y$ satisfies

$$f(\alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2) = \alpha_1 f(\mathbf{x}_1) + \alpha_2 f(\mathbf{x}_2), \quad (4.4.1)$$

then f is called a linear mapping or a linear operator. Moreover, f can also be called a linear form. Furthermore, when the mapping $f : X \rightarrow Y$ is a bijection (one-to-one and onto mapping), f is said to be an isomorphism.

For example, the derivative operator $\mathcal{D} = (\partial/\partial x_i)_{i \in \{1, \dots, d\}} : C^1(\mathbb{R}^d; \mathbb{R}) \rightarrow C(\mathbb{R}^d; \mathbb{R}^d)$ of a function $u \in C^1(\mathbb{R}^d; \mathbb{R})$ is a linear operator based on the fact that

$$\mathcal{D}(\alpha_1 u_1 + \alpha_2 u_2) = \alpha_1 \mathcal{D}(u_1) + \alpha_2 \mathcal{D}(u_2)$$

is satisfied.

Furthermore, if a linear operator f satisfies

$$\sup_{\mathbf{x} \in X \setminus \{\mathbf{0}_X\}} \frac{\|f(\mathbf{x})\|_Y}{\|\mathbf{x}\|_X} < \infty, \quad (4.4.2)$$

then f is called a bounded linear operator. In this book, the entire set of bounded linear operators from X to Y is expressed as $\mathcal{L}(X; Y)$. Moreover, if f satisfies Eq. (4.4.2), then $f : X \rightarrow Y$ is in fact continuous. This is because there exists some positive constant β and

$$\|f(\mathbf{x}) - f(\mathbf{y})\|_Y \leq \beta \|\mathbf{x} - \mathbf{y}\|_X$$

holds with respect to an arbitrary $\mathbf{x}, \mathbf{y} \in X$. Moreover, if $f : X \rightarrow Y$ is continuous, then f is bounded (cf. [104, Theorem 2.8-3, p. 104]). Here, a bounded linear operator is also called a continuous linear operator. Furthermore, the following results can be obtained with respect to the entire set of bounded linear operators, $\mathcal{L}(X; Y)$ (cf. [105, Theorem 7.6, p. 150]).

Proposition 4.4.1 (Bounded Linear Operators) *When X and Y are Banach spaces, $\mathcal{L}(X; Y)$ is a Banach space with*

$$\|f\|_{\mathcal{L}(X; Y)} = \sup_{x \in X \setminus \{0_X\}} \frac{\|f(x)\|_Y}{\|x\|_X}$$

as the norm $\|\cdot\|_{\mathcal{L}(X; Y)}$. □

Let us give an example of bounded linear operators. With n and m as natural numbers, the matrix $\mathbb{R}^{n \times m}$ with n rows and m columns is a bounded linear operator and the set of all $\mathbb{R}^{n \times m}$ can be expressed as $\mathcal{L}(\mathbb{R}^m; \mathbb{R}^n)$. When we set $y = Ax$ with respect to $x \in \mathbb{R}^m$, the norm of the matrix $A \in \mathbb{R}^{n \times m}$ is defined by

$$\|A\|_{\mathbb{R}^{n \times m}} = \|y\|_{\mathcal{L}(\mathbb{R}^m; \mathbb{R}^n)} = \sup_{x \in \mathbb{R}^m \setminus \{0_{\mathbb{R}^m}\}} \frac{\|Ax\|_{\mathbb{R}^n}}{\|x\|_{\mathbb{R}^m}}. \quad (4.4.3)$$

From this definition, when A is a positive definite real matrix ($n = m$) and the Euclidean norm $\|x\|_{\mathbb{R}^n}$ is used, the norm $\|A\|_{\mathbb{R}^{n \times n}}$ is given by the maximum eigenvalue.

4.4.2 Trace Theorem

In boundary value problems of partial differential equations, which will be looked at in detail in Chap. 5, operations for extracting values at the boundary from functions defined on a domain will be studied. Such an operation is conducted using trace operators. These operators are bounded linear operators. Let us express the outward unit normal at the boundary (Definition A.5.4) as ν and write $\partial_\nu = \nu \cdot \nabla$. In such a case, the following trace theorem can be obtained (cf. [56, Theorem 1.5.1.2, p. 37, and Theorem 1.5.1.3, p. 38])

Theorem 4.4.2 (Trace Theorem) *Let $k, l \in \mathbb{N}$, $\sigma \in (0, 1)$, $p \in (1, \infty)$, $s - 1/p = l + \sigma$ and $s \leq k + 1$. A boundary $\partial\Omega$ of $\Omega \subset \mathbb{R}^d$ is a $C^{k,1}$ class boundary when $k \geq 1$ and a Lipschitz boundary when $k = 0$. In this case, a bounded linear operator $\gamma : W^{s,p}(\Omega; \mathbb{R}) \rightarrow \prod_{i \in \{0, 1, \dots, l\}} W^{s-i-1/p, p}(\partial\Omega; \mathbb{R})$ that satisfies*

$$\gamma f = \left\{ f|_{\partial\Omega}, \partial_\nu f|_{\partial\Omega}, \dots, \partial_\nu^l f \Big|_{\partial\Omega} \right\}$$

with respect to $f \in C^{k,1}(\Omega; \mathbb{R})$ uniquely exists. This operator has a continuous right inverse operator (if $\gamma^{-1}g = f$, $\gamma f = g$ is satisfied) not dependent on p . □

The mapping γ in Theorem 4.4.2 is called a trace operator. In this book, the case $s = 1$ ($l = 0$) is assumed on the whole so it is used with the meaning $\gamma f = f|_{\partial\Omega}$.

When the dimension of a domain with respect to a function was changed from d dimensions to $d - 1$ dimensions via trace operators, the order of the derivative changed from s to $t = s - 1/p$. The reason for such a change follows from the fact that the index of differentiability remains unchanged as

$$s - \frac{d}{p} = \left(s - \frac{1}{p} \right) - \frac{d-1}{p} = t - \frac{d-1}{p}.$$

Moreover, the following results can be obtained for a function in $W_0^{s,p}(\Omega; \mathbb{R})$ (Definition 4.3.10) (cf. [56, Theorem 1.5.1.5, p. 38 and Corollary 1.5.1.6, p. 39]).

Theorem 4.4.3 (Trace Theorem with Respect to $W_0^{s,p}(\partial\Omega; \mathbb{R})$) *Let $k, l \in \mathbb{N}$, $\sigma \in (0, 1)$, $p \in (1, \infty)$, $s - 1/p = l + \sigma$ and $s \leq k + 1$. A boundary $\partial\Omega$ of $\Omega \subset \mathbb{R}^d$ is a $C^{k,1}$ class boundary when $k \geq 1$ and a Lipschitz domain when $k = 0$. Here, $f \in W_0^{s,p}(\Omega; \mathbb{R})$ is equivalent to $f \in W^{s,p}(\Omega; \mathbb{R})$ and*

$$\gamma f = \gamma \partial_\nu f = \dots = \gamma \partial_\nu^l f = 0 \quad \text{on } \partial\Omega$$

being satisfied. \square

In view of Theorem 4.4.3, the space $H_0^1(\Omega; \mathbb{R}) = W_0^{1,2}(\Omega; \mathbb{R})$ is defined as

$$H_0^1(\Omega; \mathbb{R}) = \left\{ u \in H^1(\Omega; \mathbb{R}) \mid u = 0 \text{ on } \partial\Omega \right\}.$$

This function space is used beyond Chap. 5 when considering the solution to partial differential equations using the condition for which it is zero at the boundary (homogeneous Dirichlet condition). The space $H_0^1(\Omega; \mathbb{R})$ is a linear subspace of $H^1(\Omega; \mathbb{R})$ and a real Hilbert space. In this case, the inner product and norm used are the same as for $H^1(\Omega; \mathbb{R})$.

4.4.3 Calderón Extension Theorem

In shape optimization problems with varying domain examined in Chap. 9, it is assumed that the domain in which a boundary value problem of partial differential equation is defined itself fluctuates. Therefore, the given functions used in defining the boundary value problems and the solution function need to be elements of function spaces such that they are also defined in the domain after variation. Here, the following Calderón extension theorem is used in order to extend a bounded domain Ω to \mathbb{R}^d . The bounded linear operator which existence is guaranteed in this theorem is called the extension operator (cf. [2, Theorem 5.28, p. 156]).

Theorem 4.4.4 (Calderón Extension Theorem) *Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain. For every $k \in \mathbb{N}$ and $p \in (1, \infty)$, there exists a bounded linear operator*

$$e_\Omega : W^{k,p}(\Omega; \mathbb{R}) \rightarrow W^{k,p}(\mathbb{R}^d; \mathbb{R})$$

and with respect to an arbitrary $u \in W^{k,p}(\Omega; \mathbb{R})$, we have

$$\begin{aligned} e_\Omega(u) &= u \quad \text{in } \Omega, \\ \|e_\Omega(u)\|_{W^{k,p}(\mathbb{R}^d; \mathbb{R})} &\leq c \|u\|_{W^{k,p}(\Omega; \mathbb{R})}, \end{aligned}$$

where c is a constant dependent only on k and p . \square

Note that it is $k \geq 1$ in Theorem 4.4.4.

4.4.4 Bounded Bilinear Operators

Operators with bilinearity can also be defined. Let X, Y, Z be a Banach space on K . If the mapping $f : X \times Y \rightarrow Z$ satisfies

$$\begin{aligned} f(\alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2, \mathbf{y}_1) &= \alpha_1 f(\mathbf{x}_1, \mathbf{y}_1) + \alpha_2 f(\mathbf{x}_2, \mathbf{y}_1), \\ f(\mathbf{x}_1, \alpha_1 \mathbf{y}_1 + \alpha_2 \mathbf{y}_2) &= \alpha_1 f(\mathbf{x}_1, \mathbf{y}_1) + \alpha_2 f(\mathbf{x}_1, \mathbf{y}_2) \end{aligned}$$

with respect to arbitrary $\mathbf{x}_1, \mathbf{x}_2 \in X, \mathbf{y}_1, \mathbf{y}_2 \in Y$ and $\alpha_1, \alpha_2 \in K$, then f is called a bilinear operator or a bilinear form.

Furthermore, when

$$\sup_{\mathbf{x} \in X \setminus \{\mathbf{0}_X\}, \mathbf{y} \in Y \setminus \{\mathbf{0}_Y\}} \frac{\|f(\mathbf{x}, \mathbf{y})\|_Z}{\|\mathbf{x}\|_X \|\mathbf{y}\|_Y} < \infty$$

is satisfied with respect to an arbitrary $(\mathbf{x}, \mathbf{y}) \in X \times Y$, f is called a bounded bilinear operator. In this section the entire set of bounded linear operator from $X \times Y$ to Z is expressed as $\mathcal{L}(X, Y; Z)$.

Examples of bounded bilinear operators include the kinetic energy $\kappa(u, \dot{u})$ used in Eq. (4.1.1) and elastic potential energy $\pi_I(u)$ used in Eq. (4.1.8). The expression $\kappa(u, \dot{u})$ can be written as $b(\dot{u}, \dot{u})$, where $b(u, v)$ is defined by Eq. (4.6.10), focusing on bilinearity with respect to u . Meanwhile, the functional $\pi_I(u)$ can be written as $a(u, u)$ using $a(u, v)$ defined by Eq. (4.6.17), focusing on the bilinearity with respect to u . Although these are bounded bilinear operators, they are operators with \mathbb{R} as the range. Such an operator is a functional defined more detail in the next section. Here, $b(\cdot, \cdot)$ or $a(\cdot, \cdot)$ are examples of bounded bilinear functionals.

4.4.5 *Bounded Linear Functional*

Operators with \mathbb{R} (or \mathbb{C}) as range are called functionals. In function optimization problems, the cost function is given as a mapping from a function space to real numbers. Hence, to know the properties of functionals one needs to know the properties of cost functions.

The linearity and boundedness of a functional is defined by the relationships of Eqs. (4.4.1) and (4.4.2) taking $Y = \mathbb{R}$. However, expressing them as $f(\cdot) = \langle \phi, \cdot \rangle : X \rightarrow \mathbb{R}$, a bounded linear functional is defined in the following way. Let X be a Banach space on K . If the functional $\langle \phi, \cdot \rangle : X \rightarrow \mathbb{R}$ satisfies

$$\langle \phi, \alpha_1 x_1 + \alpha_2 x_2 \rangle = \alpha_1 \langle \phi, x_1 \rangle + \alpha_2 \langle \phi, x_2 \rangle$$

with respect to arbitrary $x_1, x_2 \in X$ and $\alpha_1, \alpha_2 \in K$, then $\langle \phi, \cdot \rangle$ is called a linear functional on X . Furthermore, when

$$\sup_{x \in X \setminus \{0_X\}} \frac{|\langle \phi, x \rangle|}{\|x\|_X} < \infty$$

is satisfied, $\langle \phi, \cdot \rangle$ is called a bounded linear functional on X .

If X is a finite-dimensional vector space \mathbb{R}^d and $\phi \in \mathbb{R}^d$ is selected and fixed, a functional $\langle \phi, \cdot \rangle_{\mathbb{R}^d} : \mathbb{R}^d \rightarrow \mathbb{R}$ using the inner product is a bounded linear functional on $X = \mathbb{R}^d$.

4.4.6 *Dual Space*

If X were a finite-dimensional vector space \mathbb{R}^d , choosing a bounded linear functional $\langle \phi, \cdot \rangle_{\mathbb{R}^d}$ on $X = \mathbb{R}^d$ is equivalent to choosing an element ϕ on \mathbb{R}^d . Here the entire set \mathbb{R}^d of ϕ identified as a bounded linear functional, distinguishing it from X , is called the dual space of X and expressed as $X' = \mathbb{R}^d$. Generalizing this, the dual space can be defined as follows.

Definition 4.4.5 (Dual Space) When X is a Banach space, the entire set $\mathcal{L}(X; \mathbb{R})$ of bounded linear functionals on X is written as X' and is called the dual space of X . Moreover, $\langle \cdot, \cdot \rangle : X' \times X \rightarrow \mathbb{R}$ is also expressed as $\langle \cdot, \cdot \rangle_{X' \times X}$ and is called a dual product. \square

The dual space in Definition 4.4.5 can also be called an adjoint space.

Based on this definition, from the fact that a dual space of a Banach space is a set of all bounded linear operators $\mathcal{L}(X; \mathbb{R})$, a dual space of a Banach space is a Banach space with

$$\|\phi\|_{X'} = \sup_{x \in X \setminus \{0_X\}} \frac{|\langle \phi, x \rangle|}{\|x\|_X} \quad (4.4.4)$$

as the norm from Proposition 4.4.1.

Weak Complete and Dual Weak Complete

In discussions so far it became clear that not only the Banach spaces themselves but their dual spaces are also Banach spaces (complete norm spaces) with respect to norms such as Eq. (4.4.4). In other words, it indicates that Cauchy series measured by the various norms always converge. In the case of a finite-dimensional vector space, its dual space is also the same finite-dimensional vector space. Hence, convergence could be measured using the same norm. However, the definition of the norm generally differs for a Banach space and its dual space. For this reason, convergence other than that using the norm can be defined too. Here, let us define convergence using the dual product.

Definition 4.4.6 (Weak Convergence) Let X be a Banach space and X' its dual space. When an infinite sequence of points $\{x_n\}_{n \in \mathbb{N}} \in X$ with respect to an arbitrary $\phi \in X'$ satisfies

$$\lim_{n,m \rightarrow \infty} \langle \phi, x_n - x_m \rangle = 0,$$

$\{x_n\}_{n \in \mathbb{N}}$ is called a weak Cauchy sequence. The convergence of a weak Cauchy sequence is called a weak convergence and is written as $x_n \rightarrow x$ weakly in X . When any weak Cauchy sequence of X converges within X , then X is said to be weak complete. Furthermore, when an arbitrary infinite sequence of points of a subset V of weakly complete X includes an infinite subsequence that weakly converges within V , then V is referred to as weak compact. \square

By contrast, convergence relating to the norm is called a strong convergence and written as $x_n \rightarrow x$ strongly in X . Also, if the purposes of a Banach space and its dual space are reversed, a definition of another convergence is possible.

Definition 4.4.7 (Dual Weak Convergence) Let X be a Banach space and X' be its dual space. When the infinite sequence of points $\{\phi_n\}_{n \in \mathbb{N}} \in X'$ satisfies

$$\lim_{n,m \rightarrow \infty} \langle \phi_n - \phi_m, x \rangle = 0$$

with respect to an arbitrary $x \in X$, then $\{\phi_n\}_{n \in \mathbb{N}} \in X'$ is called a dual weak Cauchy sequence. Convergence of a dual weak Cauchy sequence is called a dual weak convergence and is expressed as $\phi_n \rightarrow \phi$ \ast -weakly in X' . When any of the dual weak Cauchy sequence in X' converges to a point within X' , X' is said to be dual weak complete. Furthermore, if an arbitrary infinite sequence of points of the subset V' of dual weak complete X' includes an infinite subsequence that is dual weakly convergent in V' , then V' is said to be dual weak compact. \square

As shown later, the Fréchet derivative of the cost function in a function optimization problem is defined using the dual product. Weak completeness and dual weak completeness are qualities which are necessary when seeking the minimum point using the Fréchet derivative of the cost function.

Let us consider the conditions which guarantee that a Banach space has weak completeness. For this purpose, a reflexive Banach space is defined as follows.

Definition 4.4.8 (Reflexive Banach Space) Let X be a Banach space. Let X' and $X'' = (X')'$ be the dual space and second dual space of X , respectively. If an evaluation map $\tau : X \rightarrow X''$ ($x \in X$ generates a scalar with respect to $f \in X'$) such that

$$\langle f, \tau(x) \rangle_{X' \times X''} = \langle f, x \rangle_{X' \times X}$$

holds with respect to all $(x, f) \in X \times X'$ is a one-to-one and onto mapping, then X is called a reflexive Banach space. \square

The following results can be obtained with respect to a reflexive Banach space (cf. [105, Theorem 8.33, p. 193]).

Proposition 4.4.9 (Weak Complete) *A reflexive Banach space is weakly complete.* \square

From Proposition 4.4.9, if they are in a reflexive Banach space, weak Cauchy series or dual weak Cauchy series are guaranteed to converge to an element within that space. A Sobolev space is a Banach space (Proposition 4.3.11), but regarding reflexivity of a Sobolev space, the following result can be obtained (cf. [116, Theorem 2.25, p. 41]).

Proposition 4.4.10 (Separability and Reflexivity of Sobolev Spaces) *Let $\Omega \subset \mathbb{R}^d$. If $p \in (1, \infty)$ (note it is not $[1, \infty]$) and $k \in \mathbb{N} \cup \{0\}$, then $W^{k,p}(\Omega; \mathbb{R})$ and $W_0^{k,p}(\Omega; \mathbb{R})$ are reflexive.* \square

Since $H^k(\Omega; \mathbb{R})$ is a Sobolev space when $p = 2$, it is a reflexive Banach space. Therefore, from Proposition 4.4.10, $H^k(\Omega; \mathbb{R})$ is weakly complete. On the other hand, from the fact that $L^1(\Omega; \mathbb{R})$ or $L^\infty(\Omega; \mathbb{R})$ are not reflexive, they are not weakly complete.

Moreover, the following is known about the compactness of Banach spaces. A unit sphere in infinite-dimensional space is not compact. This is because, selecting an infinite number of basic vectors such as $(1, 0, 0, \dots), (0, 1, 0, \dots), \dots$, an infinite

sequence of points which does not include a Cauchy sequence can be constructed (cf. [183, Subsection 1.2.1, p. 15]). However, the following result is known (cf. [105, Theorem 8.36, p. 194]).

Proposition 4.4.11 (Weak Compact) *A closed unit sphere in a reflexive Banach space is a weak compact.* \square

Dual Space of Sobolev Space

With respect to the dual space of a Sobolev space, a clear result can be obtained. Firstly, with respect to index p of $L^p(\Omega; \mathbb{R})$, a duality index q is defined as follows.

Definition 4.4.12 (Duality Index) For $p \in [1, \infty)$, a constant $q \in [1, \infty]$ satisfying

$$\frac{1}{q} + \frac{1}{p} = 1$$

is called a duality index. Moreover, with respect to $L^p(\Omega; \mathbb{R})$, $L^q(\Omega; \mathbb{R})$ is called a dual space of $L^p(\Omega; \mathbb{R})$ and is expressed as $(L^p(\Omega; \mathbb{R}))'$. \square

Using the duality index, let us see how the dual space of Sobolev space $W^{k,p}(\Omega; \mathbb{R})$ for $k \geq 1$ is defined. Firstly, let $\Omega = (0, 1)$ and consider the dual space $(H^1((0, 1); \mathbb{R}))'$ of $H^1((0, 1); \mathbb{R})$. When an arbitrary $f \in (H^1((0, 1); \mathbb{R}))'$ is selected, f becomes a bounded linear functional with respect to an arbitrary $v \in H^1((0, 1); \mathbb{R})$. According to the Riesz representation theorem (Theorem 4.4.17) which will be shown later, there exists a unique $u \in H^1((0, 1); \mathbb{R})$ which satisfies

$$\langle f, v \rangle = (u, v)_{H^1((0, 1); \mathbb{R})} = \int_0^1 (uv + u'v') \, dx \quad (4.4.5)$$

with respect to an arbitrary $v \in H^1((0, 1); \mathbb{R})$. Hence, $f_0, f_1 \in L^2((0, 1); \mathbb{R})$ exist which satisfy

\langle f, v \rangle = f(v) = \int_0^1 (f_0 v + f_1 v') \, dx. \quad (4.4.6)

The norm of f is defined by

$$\|f\|_{(H^1((0, 1); \mathbb{R}))'} = \sup_{v \in H^1((0, 1); \mathbb{R})} \frac{|\langle f, v \rangle|}{\|v\|_{H^1((0, 1); \mathbb{R})}}.$$

Here, since

$$\begin{aligned} |\langle f, v \rangle| &= \left| \int_0^1 (f_0 v + f_1 v') dx \right| \\ &\leq (\|f_0\|_{L^2((0,1);\mathbb{R})} + \|f_1\|_{L^2((0,1);\mathbb{R})}) \|v\|_{H^1((0,1);\mathbb{R})} \end{aligned}$$

is established from the Schwarz inequality (see Theorem A.9.1), we expect that

$$\begin{aligned} \|f\|_{(H^1((0,1);\mathbb{R}))'} &= \inf_{f_0, f_1 \in L^2((0,1);\mathbb{R})} \left\{ \|f_0\|_{L^2((0,1);\mathbb{R})} + \|f_1\|_{L^2((0,1);\mathbb{R})} \mid \text{Eq. (4.4.6)} \right\} \end{aligned}$$

can also be established.

Generalizing this leads to the following result (cf. [2, Theorem 3.8 and Theorem 3.9, p. 62], [116, Theorem 2.20 and Theorem 2.21, p. 38]).

Proposition 4.4.13 (Dual Space of $W^{k,p}(\Omega; \mathbb{R})$) *Let $\Omega \subset \mathbb{R}^2$. Let $k \in \mathbb{N} \cup \{0\}$ and $p \in [1, \infty)$. Also, let q be a dual index with respect to p and $(W^{k,p}(\Omega; \mathbb{R}))'$ be the dual space of $W^{k,p}(\Omega; \mathbb{R})$. For an arbitrary $f \in (W^{k,p}(\Omega; \mathbb{R}))'$, there exists $f_\beta \in L^q(\Omega; \mathbb{R})$ which satisfies*

$$f(v) = \sum_{|\beta| \leq k} \int_{\Omega} \nabla^\beta v f_\beta dx \quad (4.4.7)$$

for all $v \in W^{k,p}(\Omega; \mathbb{R})$. Furthermore,

$$\|f\|_{(W^{k,p}(\Omega; \mathbb{R}))'} = \inf_{f_\beta \in L^q(\Omega; \mathbb{R}), |\beta| \leq k} \left\{ \sum_{|\beta| \leq k} \|f_\beta\|_{L^q(\Omega; \mathbb{R})} \mid \text{Eq. (4.4.7)} \right\}$$

holds. □

In Proposition 4.4.13, the case of $p = \infty$ is removed. An explanation of the dual space of $L^\infty(\Omega; \mathbb{R})$ can be found in [184, Example 5, p. 118].

The following results can be obtained with respect to the dual space of the Sobolev space $W_0^{k,p}(\Omega; \mathbb{R})$. Here, f_β is an element of $L^q(\Omega; \mathbb{R})$ shown to exist by Proposition 4.4.13. Functions g and g_β are taken to be elements of $(C_0^\infty(\Omega; \mathbb{R}))'$ satisfying

$$g(\phi) = \sum_{|\beta| \leq k} (-1)^{|\beta|} \nabla^\beta g_\beta(\phi), \quad g_\beta(\phi) = \int_{\Omega} \phi f_\beta dx \quad \text{for } |\beta| \leq k \quad (4.4.8)$$

with respect to an arbitrary $\phi \in C_0^\infty(\Omega; \mathbb{R})$. Here, from the fact that

$$\nabla^\beta g_\beta(\phi) = (-1)^{|\beta|} \int_\Omega \nabla^\beta \phi f_\beta \, dx$$

holds with respect to an arbitrary $\phi \in C_0^\infty(\Omega; \mathbb{R})$, the identity

$$g(\phi) = \sum_{|\beta| \leq k} g_\beta(\nabla^\beta \phi) = f(\phi)$$

can be obtained. Here, Eq. (4.4.7) was used. This result shows that $f \in (W^{k,p}(\Omega; \mathbb{R}))'$ is a function which expands $C_0^\infty(\Omega; \mathbb{R})$ on the Schwartz distribution $g \in (C_0^\infty(\Omega; \mathbb{R}))'$ to $W^{k,p}(\Omega; \mathbb{R})$. As a result, it should be noted that with respect to embedding relationships of $C_0^\infty(\Omega; \mathbb{R}) \subset W_0^{k,p}(\Omega; \mathbb{R}) \subset W^{k,p}(\Omega; \mathbb{R})$, the embedding relationships of their dual spaces become $(W^{k,p}(\Omega; \mathbb{R}))' \subset (W_0^{k,p}(\Omega; \mathbb{R}))' \subset (C_0^\infty(\Omega; \mathbb{R}))'$ (Practice 4.4).

These relationships give the following results (cf. [2, Theorem 3.12, p. 64], [116, Theorem 2.3, p. 40]. With respect to $(H_0^1(\Omega; \mathbb{R}))'$ [47, Theorem 1, p. 283], [183, Example 3.4, p. 80]).

Proposition 4.4.14 (Dual Space of $W_0^{k,p}(\Omega; \mathbb{R})$) *Let $\Omega \subset \mathbb{R}^d$, $k \in \mathbb{N}$ and $p \in [0, \infty)$. Also, let v_β be an element of $L^q(\Omega; \mathbb{R})$ which has been identified by Proposition 4.4.13. Let $(W_0^{k,p}(\Omega; \mathbb{R}))'$ be the dual space of $W_0^{k,p}(\Omega; \mathbb{R})$. In this case, $g \in (W_0^{k,p}(\Omega; \mathbb{R}))'$ is uniquely determined in the sense of the Schwartz distribution $g \in (C_0^\infty(\Omega; \mathbb{R}))'$ by Eq. (4.4.8). Furthermore,*

$$\|g\|_{(W_0^{k,p}(\Omega; \mathbb{R}))'} = \inf_{g_\beta \in L^q(\Omega; \mathbb{R}), |\beta| \leq k} \left\{ \sum_{|\beta| \leq k} \|g_\beta\|_{L^q(\Omega; \mathbb{R})} \mid \text{Eq. (4.4.8)} \right\}$$

is established. \square

From Proposition 4.4.14, the element of $(W_0^{k,p}(\Omega; \mathbb{R}))'$ is a function for which the k -th integral (note that it is not a differential) is q -th order integrable. From this fact, $(W_0^{k,p}(\Omega; \mathbb{R}))'$ can also be expressed as $W^{-k,q}(\Omega; \mathbb{R})$. Moreover, the case $k = 0$ is excluded in Proposition 4.4.14 because $W_0^{0,p}(\Omega; \mathbb{R}) = L^p(\Omega; \mathbb{R})$ was defined in Definition 4.3.10.

4.4.7 Rellich–Kondrachov Compact Embedding Theorem

The result which rewrites the embedding relationship of Sobolev space given by Sobolev embedding theorem (Theorem 4.3.14) as an embedding relationship with compactness is called the Rellich–Kondrachov compact embedding theorem. Here, a Banach space X compactly embedded in a Banach space Y is defined by:

- (1) $\|\phi\|_Y \leq c \|\phi\|_X$ being established with respect to an arbitrary $\phi \in Y$ and
- (2) an arbitrary bounded infinite sequence of points in X including a subsequence converging to within Y with the norm of Y (X is relative compact)).

Here, it is written as $X \Subset Y$. Therefore, if X is weakly complete, X is complete with the norm $\|\cdot\|_Y$ of Y .

The following result has been obtained (cf. [2, Theorem 6.3, p. 168], [116, Chap. 7, p. 153]).

Theorem 4.4.15 (Rellich–Kondrachov Compact Embedding Theorem) *Let $\Omega \subset \mathbb{R}^d$, $k \in \mathbb{N}$, $j \in \mathbb{N} \cup \{0\}$ and $p \in [1, \infty)$. In this case:*

- (1) *if $k - d/p < 0$, with $p^* = d / \{(d/p) - k\}$, then the embedding*

$$W^{k+j,p}(\Omega; \mathbb{R}) \Subset W^{j,q}(\Omega; \mathbb{R}) \quad \text{for } q \in [1, p^*], \quad (4.4.9)$$

- (2) *if $k - d/p = 0$, then the embedding*

$$W^{k+j,p}(\Omega; \mathbb{R}) \Subset W^{j,q}(\Omega; \mathbb{R}) \quad \text{for } q \in [1, \infty), \quad (4.4.10)$$

- (3) *if $k - d/p = j + \sigma > 0$ ($\sigma \in (0, 1)$), or when $k = d$ and $p = 1$, then the embedding*

$$W^{k+j,p}(\Omega; \mathbb{R}) \Subset W^{j,q}(\Omega; \mathbb{R}) \quad \text{for } q \in [p, \infty) \quad (4.4.11)$$

holds. Furthermore, if Ω is a Lipschitz domain, then

- (4) *for $k - d/p = j + \sigma > 0$ ($\sigma \in (0, 1)$), or when $k = d$ and $p = 1$, we have*

$$W^{k+j,p}(\Omega; \mathbb{R}) \Subset C^{j,\lambda}(\bar{\Omega}; \mathbb{R}) \quad \text{for } \lambda \in (0, \sigma]. \quad (4.4.12)$$

□

If the Sobolev embedding theorem (Theorem 4.3.14) and Rellich–Kondrachov compact embedding theorem (Theorem 4.4.15) are compared, the condition $q \in [p, p^*]$ in Eq. (4.3.16) is different in that it is $q \in [1, p^*]$ in Eq. (4.4.9).

Based on Theorem 4.4.15, the following result can be obtained in relation to the completeness of $H^k(\Omega; \mathbb{R})$.

Proposition 4.4.16 (Completeness of $H^k(\Omega; \mathbb{R})$ in $L^2(\Omega; \mathbb{R})$) *An arbitrary infinite sequence of points of $H^k(\Omega; \mathbb{R})$ ($k \in \{1, 2, \dots\}$) includes an infinite subsequence which strongly converges in $H^{k-1}(\Omega; \mathbb{R})$ using $\|\cdot\|_{H^{k-1}(\Omega; \mathbb{R})}$.* \square

The result of Proposition 4.4.16 relates in the following way to the optimum design problem defined in function space shown in Chap. 7 and beyond. In shape optimization problems shown in Chaps. 8 and 9, we assume that the linear space X which contains the design variables is defined with $H^1(\Omega; \mathbb{R})$ or $H^1(\Omega; \mathbb{R}^d)$, and the admissible set \mathcal{D} is defined with $H^2(\Omega; \mathbb{R})$ or $H^2(\Omega; \mathbb{R}^d)$. Here if the trial points are updated using the gradient method or the Newton method, such a sequence of points can be thought to be an infinite sequence of points on \mathcal{D} . When considering its convergence, Proposition 4.4.16 shows the fact that the existence of a strongly convergent subsequence using the norm of X is guaranteed.

4.4.8 Riesz Representation Theorem

Before we end this section (Sect. 4.4), let us discuss the Riesz representation theorem which is used when showing the unique existence of the solution to boundary value problem of elliptic partial differential equation (Definition A.7.1) (cf. [2, Theorem 1.12, p. 6], [183, Theorem 3.6, p. 79]).

Theorem 4.4.17 (Riesz Representation Theorem) *Let X be a Hilbert space, X' be the dual space of X , $(\cdot, \cdot)_X$ the inner product on X and $\langle \cdot, \cdot \rangle_{X' \times X}$ be the dual product. For any $\phi \in X'$, there exist some unique $x \in X$ such that*

$$\langle \phi, y \rangle_{X' \times X} = (x, y)_X, \quad \|\phi\|_{X'} = \|x\|_X$$

holds for every $y \in X$. Moreover, there exists an isomorphism $\tau : X' \rightarrow X$ which satisfies

$$\langle \phi, y \rangle_{X' \times X} = (\tau\phi, y)_X, \quad \|\tau\|_{\mathcal{L}(X'; X)} = 1. \quad \square$$

When X is a finite-dimensional vector space \mathbb{R}^d , the dual product matches the inner product and $X' = X$ with τ becoming the identity mapping.

Using the isomorphism τ whose existence was shown via Theorem 4.4.17, the inner product for the dual space X' of a Hilbert space X can be defined as $(\phi, \varphi)_{X'} = (\tau\phi, \tau\varphi)_X$. If this inner product is used, X' becomes a Hilbert space too.

The Riesz representation theorem will be referred to as the Lax–Milgram theorem (Theorem 5.2.4) in Chap. 5 and will be used in Chap. 5 as well as in Chap. 7 and beyond.

4.5 Generalized Derivatives

In Sect. 4.4, bounded linear operators and bounded linear functionals in Banach space were defined and we saw how the set of all bounded linear functionals becomes a dual space. In this section, this relationship is used to show the definition of a derivative with respect to operators or functionals. Here, the definition of a Gâteaux derivative, which is also called a directional derivative, is shown first, then the definition of the Fréchet derivative, which may define gradients, is provided. In particular, a gradient in the Fréchet derivative of a functional is defined as an element of dual space with respect to a Banach space of a variation vector. This relationship is an essential relationship when applying optimization theorems using derivatives of cost functions, as seen in Chap. 2, or gradient methods, looked at in Chap. 3, to function optimization problems.

4.5.1 Gâteaux Derivative

Firstly, let us start by looking at the Gâteaux derivative defined in loose conditions. In this section, X and Y are assumed to be Banach spaces and functions (operators) are seen to be given by mappings from X to Y .

Definition 4.5.1 (k -th Order Gâteaux Derivative) Let X and Y be Banach spaces on \mathbb{R} . With respect to a neighborhood (open set) $B \subset X$ of $x \in X$, suppose $f : B \rightarrow Y$ is defined. Choose $y \in X$ to be a variation vector and fix it. Let $k \in \mathbb{N}$. When the mapping $\epsilon \mapsto f(x + \epsilon y)$ is an element of $C^k(\mathbb{R}; Y)$ with respect to an arbitrary $\epsilon \in \mathbb{R}$,

$$f^{(k)}(x)[y] = \frac{d^k}{d\epsilon^k} f(x + \epsilon y) \Big|_{\epsilon=0}$$

is called the k -th order Gâteaux derivative of f on x in the direction y . □

There are times when an alternative definition of a Gâteaux derivative to Definition 4.5.1 is used. When a beneficial result can be derived from these definitions, those definitions should be used. However, in this book, we will not step out any further since we do not go on to proper discussions using Gâteaux derivatives.

Let us apply the definition of a Gâteaux derivative with respect to f of variational Problem 4.1.1.

Exercise 4.5.2 (Gâteaux Derivative of an Extended Action Integral) Show the Gâteaux derivative and second-order Gâteaux derivative of an extended action integral

$$f(u) = \int_0^{t_T} l(u) \, dt - m\beta u(t_T)$$

defined in Problem 4.1.1. \square

Answer In Problem 4.1.1, the set of functions u satisfying $u(0) = \alpha$ is expressed as U and the set of functions v satisfying $v(0) = 0$ is expressed as V . The set U is not a linear space. However, V is a linear space. Here, the Banach space X in Definition 4.5.1 is chosen to be V and $X = \{v \in H^1((0, t_T); \mathbb{R}) \mid v(0) = 0\}$ (explained in Sect. 4.6.1). The problem of finding $u \in U$ is equivalent to seeking $u - u_0 \in V$ after selecting and fixing a $u_0 \in U$. Moreover, the Banach space Y is taken as the range \mathbb{R} .

Under these assumptions, let the fixed variation vector be $v \in X$ and seek the Gâteaux derivative in this direction. From Definition 4.5.1, we set

$$f(u + \epsilon v) = \int_0^t \left\{ \frac{1}{2}m(\dot{u} + \epsilon \dot{v})^2 - \frac{1}{2}k(u + \epsilon v)^2 + p(u + \epsilon v) \right\} dt - m\beta(u(t) + \epsilon v(t))$$

for an arbitrary $\epsilon \in \mathbb{R}$. Hence, the Gâteaux derivative is computed as follows:

$$\begin{aligned} f^{(1)}(u)[v] &= f'(u)[v] = \frac{df}{d\epsilon} \Big|_{\epsilon=0} \\ &= \left[\int_0^t \{m(\dot{u} + \epsilon \dot{v}) \dot{v} - k(u + \epsilon v)v + pv\} dt - m\beta v(t) \right] \Big|_{\epsilon=0} \\ &= \int_0^t (m\dot{u}\dot{v} - kuv + pv) dt - m\beta v(t). \end{aligned}$$

This equation matches $f'(u)[v]$ in Eq. (4.1.4). However, $f'(u)[v]$ in Eq. (4.1.4) was shown with respect to an arbitrary $v \in X$. Moreover, a second-order Gâteaux derivative becomes

$$f^{(2)}(u)[v] = f''(u)[v] = \frac{d^2f}{d\epsilon^2} \Big|_{\epsilon=0} = \int_0^t (m\dot{v}^2 - kv^2) dt.$$

This equation is the same as Eq. (4.1.4). Even in this case it is different from Eq. (4.1.4) in that $v \in X$ is fixed. \square

Next, let us look at an example for which Gâteaux differentiation is possible but Fréchet differentiation shown in the next section is not possible.

Exercise 4.5.3 (Example Which Is Only Gâteaux Differentiable) Find the Gâteaux derivative of the following function in two-dimensional space:

$$f(\mathbf{x}) = \begin{cases} \frac{x_1^3}{x_1^2 + x_2^2} & \text{for } \mathbf{x} \neq \mathbf{0}_{\mathbb{R}^2}, \\ 0 & \text{for } \mathbf{x} = \mathbf{0}_{\mathbb{R}^2} \end{cases}$$

at $\mathbf{x} = \mathbf{0}_{\mathbb{R}^2}$. □

Answer The Gâteaux derivative of $f(\mathbf{x})$ at $\mathbf{x} = \mathbf{0}_{\mathbb{R}^2}$ is given by

$$f'(\mathbf{0}_{\mathbb{R}^2})[\mathbf{y}] = \begin{cases} \frac{y_1^3}{y_1^2 + y_2^2}, & \text{for } \mathbf{y} \neq \mathbf{0}_{\mathbb{R}^2}, \\ 0, & \text{for } \mathbf{y} = \mathbf{0}_{\mathbb{R}^2}. \end{cases}$$

In view of the above expression, $f'(\mathbf{0}_{\mathbb{R}^2})[\mathbf{y}]$ is continuous with respect to \mathbf{y} but is non-linear. □

4.5.2 Fréchet Derivative

The Gâteaux derivative was defined as a derivative with respect to a variation vector once its direction was specified. However, a derivative that is required in this book is a derivative which can define a gradient. In other words, it is a derivative of a functional which may be given as a dual product of an arbitrary variation vector and gradient. The Fréchet derivative shown next is a definition of a derivative generalized as a derivative of operators between Banach spaces and not only limited to functionals.

Definition 4.5.4 (k -th Order Fréchet Derivative) Let X and Y be Banach spaces on \mathbb{R} . With respect to a neighborhood $B \subset X$ of $\mathbf{x} \in X$, suppose $f : B \rightarrow Y$ is defined. If there exists a bounded linear operator $f'(\mathbf{x})[\cdot] \in \mathcal{L}(X; Y)$ satisfying

$$\lim_{\|\mathbf{y}_1\|_X \rightarrow 0} \frac{\|f(\mathbf{x} + \mathbf{y}_1) - f(\mathbf{x}) - f'(\mathbf{x})[\mathbf{y}_1]\|_Y}{\|\mathbf{y}_1\|_X} = 0 \quad (4.5.1)$$

with respect to an arbitrary variation vector $\mathbf{y}_1 \in X$, then $f'(\mathbf{x})[\mathbf{y}_1]$ is called the Fréchet derivative of f at \mathbf{x} . Moreover, when there exists $f'(\mathbf{x})[\mathbf{y}_1]$ with respect to all $\mathbf{x} \in B$ and those are in $C(B; \mathcal{L}(X; Y))$, it is expressed as $f \in C^1(B; Y)$.

Furthermore, if there exists a functional $f''(\mathbf{x})[\mathbf{y}_1, \cdot] \in \mathcal{L}(X; \mathcal{L}(X; Y))$ which satisfies

$$\lim_{\|\mathbf{y}_2\|_X \rightarrow 0_X} \frac{\|f'(\mathbf{x} + \mathbf{y}_2)[\mathbf{y}_1] - f'(\mathbf{x})[\mathbf{y}_1] - f''(\mathbf{x})[\mathbf{y}_1, \mathbf{y}_2]\|_Y}{\|\mathbf{y}_2\|_X} = 0$$

with respect to an arbitrary $\mathbf{y}_2 \in X$, then $f''(\mathbf{x})[\mathbf{y}_1, \mathbf{y}_2]$ is called a second-order Fréchet derivative of f at \mathbf{x} . The space $\mathcal{L}(X; \mathcal{L}(X; Y))$ is expressed as $\mathcal{L}^2(X \times X; Y)$. In addition, with respect to all $\mathbf{x} \in B$, if there exists a second-order Fréchet derivative and $f''(\mathbf{x})[\cdot, \cdot] \in C(B; \mathcal{L}(X; \mathcal{L}(X; Y)))$, then we find that $f \in C^2(B; Y)$. Similarly, a $k \in \{3, 4, \dots\}$ -th order Fréchet derivative $f^{(k)}$ can be defined and is expressed as $f \in C^k(B; Y)$. \square

Equation (4.5.1) used in Definition 4.5.4 can be expressed using a Taylor expansion such as

$$f(\mathbf{x} + \mathbf{y}_1) = f(\mathbf{x}) + f'(\mathbf{x})[\mathbf{y}_1] + o(\|\mathbf{y}_1\|_X).$$

Here, $o(\|\mathbf{y}_1\|_X)$ is called the Bachmann–Landau small- o symbol and it is assumed that the limit

$$\lim_{\|\mathbf{y}_1\|_X \rightarrow 0} \frac{o(\|\mathbf{y}_1\|_X)}{\|\mathbf{y}_1\|_X} = \mathbf{0}_Y$$

holds.

Moreover, in Definition 4.5.4, if $Y = \mathbb{R}$, $f : B \rightarrow \mathbb{R}$ becomes a functional. Here, the Fréchet derivative of f can be written as

$$f'(\mathbf{x})[\mathbf{y}] = \langle \mathbf{g}, \mathbf{y} \rangle_{X' \times X}. \quad (4.5.2)$$

In this case, $\mathbf{g} \in X'$ is called a gradient. The second-order Fréchet derivative of f is written as

$$f''(\mathbf{x})[\mathbf{y}_1, \mathbf{y}_2] = h(\mathbf{x})[\mathbf{y}_1, \mathbf{y}_2], \quad (4.5.3)$$

where $h(\mathbf{x}) \in \mathcal{L}^2(X \times X; \mathbb{R})$ is called the Hesse form of f at \mathbf{x} .

Let us review the first and second variation of f in the variational Problem 4.1.1 using the definition of a Fréchet derivative in the following exercise.

Exercise 4.5.5 (Fréchet Derivative of Extended Action Integral) Obtain the Fréchet derivative and second-order Fréchet derivative of the extended action integral

$$f(u) = \int_0^{t_T} l(u) \, dt - m\beta u(t_T)$$

defined in Problem 4.1.1, where $u(0) = \alpha$. \square

Answer Let $X = \{v \in H^1((0, t_T); \mathbb{R}) \mid v(0) = 0\}$, $Y = \mathbb{R}$ and $f'(u)[\cdot]$ in Eq. (4.1.5) be in $\mathcal{L}(X; Y)$ (explained in Sect. 4.6.1). Therefore, $f'(u)[v_1]$ can be viewed as a Fréchet derivative and it can be expressed as $f'(u)[v_1] = \langle g, v_1 \rangle_{X' \times X}$. In this case, the gradient is given by $g \in X'$.

Next, when v_1 is fixed in $f'(u)[v_1]$ and an arbitrary variation $v_2 \in X$ is added to u , we have

$$\begin{aligned} f'(u + v_2)[v_1] &= \int_0^{t_T} \{m(\dot{u} + \dot{v}_2)\dot{v}_1 - k(u + v_2)v_1 + p v_1\} dt - m\beta v_1(t_T) \\ &= f'(u)[v_1] + \int_0^{t_T} (m\dot{v}_1\dot{v}_2 - k v_1 v_2) dt \\ &= f'(u)[v_1] + f''(u)[v_1, v_2]. \end{aligned}$$

Here, it follows that $f''(u)[v_1, v_2]$ is the same as $f''(u)[v, v]$ of Eq. (4.1.4). \square

The gradient $g(x)$ defined in Eq. (4.5.2) is an element of the dual space X' . So, its norm is given by

$$\begin{aligned} \|g(x)\|_{X'} &= \sup_{y \in X \setminus \{0_X\}} \frac{|\langle g(x), y \rangle_{X' \times X}|}{\|y\|_X} \\ &= \sup_{y \in X, \|y\|_X=1} |\langle g(x), y \rangle_{X' \times X}| \end{aligned} \quad (4.5.4)$$

with respect to the definition of norm for a bounded linear functional (see Fig. 3.1).

4.6 Function Spaces in Variational Principles

Various function spaces were defined in Sect. 4.3 and it was shown that they satisfy the requirements of a Banach space and Hilbert space. In Sect. 4.4, an operator was defined as a mapping from between Banach spaces and in Sect. 4.5, the derivative of operators was defined. Among them, we saw that the Fréchet derivative of a functional can be defined by a dual product of the function space containing the variable of the functional and the dual space containing the gradient. In this section, the contents seen from Sects. 4.2 to 4.5 are used to review variational principles and optimization control problems shown in Sect. 4.1.

4.6.1 Hamilton's Principle

Summarizing the definitions used in the expanded Hamilton's principle (Problem 4.1.1) we obtain the following. The set of functions $u : (0, t_T) \rightarrow \mathbb{R}$ satisfying $u(0) = \alpha$ was expressed as U . Moreover, the set of functions $v : (0, t_T) \rightarrow \mathbb{R}$ satisfying $v(0) = 0$ was denoted as V . A functional expressing extended action integral with respect to $u \in U$ was set as

$$f(u) = \int_0^{t_T} \left(\frac{1}{2} m \dot{u}^2 - \frac{1}{2} k u^2 + pu \right) dt - m\beta u(t_T). \quad (4.6.1)$$

Moreover, a stationary condition of $f(u + v)$ with respect to an arbitrary $v \in V$ satisfying $v(0) = 0$ was obtained:

$$f'(u)[v] = \int_0^{t_T} (m\dot{u}\dot{v} - kuv + pv) dt - m\beta v(t_T) = 0 \quad (4.6.2)$$

during the calculation of Eq. (4.1.3).

In order for integrals of Eqs. (4.6.1) and (4.6.2) to have meaning, there is a need to clarify the function spaces with respect to u , v and p . We do this as follows.

Firstly, let us think about what U and V were. A element $u \in U$ has to satisfy $u(0) = \alpha$. Hence, let us set

$$U = \left\{ u \in H^1((0, t_T); \mathbb{R}) \mid u(0) = \alpha \right\}. \quad (4.6.3)$$

On the other hand, since there was a need for $v \in V$ to satisfy the boundary condition of a homogeneous form $v(0) = 0$, we let

$$V = \left\{ v \in H^1((0, t_T); \mathbb{R}) \mid v(0) = 0 \right\}. \quad (4.6.4)$$

Here, U is not a linear space because of the non-homogeneous boundary condition $u(0) \neq 0$. On the other hand, V is a linear space. Moreover, U is an affine subspace (Definition 4.2.7) with respect to V . In fact, if an element u_0 of $H^1((0, t_T); \mathbb{R})$ which satisfies $u(0) = \alpha$ is chosen and fixed, U is equivalent to $V(u_0)$. Let us set $\tilde{u} = u - u_0$ and let it be an element of V .

When V and \tilde{u} are set in this way, and if Minkowski's inequality (Theorem A.9.2) and Hölder's inequality (Theorem A.9.1) are applied to the integral of $\dot{u}\dot{v}$, which is on the right-hand side of Eq. (4.6.2), then one obtains

$$\begin{aligned} \int_0^{t_T} \dot{u}\dot{v} dt &\leq \left\| \dot{\tilde{u}}\dot{v} \right\|_{L^1((0, t_T); \mathbb{R})} + \left\| \dot{u}_0\dot{v} \right\|_{L^1((0, t_T); \mathbb{R})} \\ &\leq \left\| \dot{\tilde{u}} \right\|_{L^2((0, t_T); \mathbb{R})} \left\| \dot{v} \right\|_{L^2((0, t_T); \mathbb{R})} + \left\| \dot{u}_0 \right\|_{L^2((0, t_T); \mathbb{R})} \left\| \dot{v} \right\|_{L^2((0, t_T); \mathbb{R})}. \end{aligned} \quad (4.6.5)$$

Furthermore, if the Poincaré inequality system (Corollary A.9.4) is used for the right-hand side of Eq. (4.6.5), the inequality

$$\int_0^{t_T} \dot{u} \dot{v} dt \leq \|\tilde{u}\|_V \|v\|_V + \|u_0\|_{H^1((0, t_T); \mathbb{R})} \|v\|_V \quad (4.6.6)$$

is established. This confirms that if u_0 is an element of $H^1((0, t_T); \mathbb{R})$ and \tilde{u} and v are elements of V , the right-hand side of Eq. (4.6.6) is bounded. Moreover, the integral of \dot{u}^2 and u^2 on the right-hand side of Eq. (4.6.1) as well as the integral of uv on the right-hand side of Eq. (4.6.2) are also bounded.

The fact that the boundary values $u(t_T)$ and $v(t_T)$ of u and v which appear in Eqs. (4.6.1) and (4.6.2) are defined can be verified as follows. From the Sobolev embedding theorem (Theorem 4.3.14), the inclusion $H^1((0, t_T); \mathbb{R}) \subset C^{0,1/2}([0, t_T]; \mathbb{R})$ is established. Here, u and v are continuous functions and boundary values are set (trace is taken).

Meanwhile, for the integrals of pu and pv on the right-hand side of Eqs. (4.6.1) and (4.6.2) to be bounded, it suffices to show if p is an element of $L^2((0, t_T); \mathbb{R})$. This is because the inequality

$$\int_0^{t_T} p v dt \leq \|p\|_{L^2((0, t_T); \mathbb{R})} \|v\|_V \quad (4.6.7)$$

in fact holds.

In the case that $u - u_0 \in V$, $v \in V$ and $p \in L^2((0, t_T); \mathbb{R})$, the integral $f'(u)[v]$ of Eq. (4.6.2) makes sense and $f'(u)[v]$ is seen as a bounded linear functional with respect to $v \in V$. Here, we can write

$$f'(u)[v] = \langle g, v \rangle_{V' \times V}, \quad (4.6.8)$$

where g is an element of the dual space V' of V and is called the gradient of f . The fact that Eq. (4.6.8) becomes zero with respect to an arbitrary $v \in V$ is equivalent to the equation of motion (Eq. (4.1.6)) and the terminal condition for velocity (Eq. (4.1.7)) being established.

In order to use the above ideas in Chap. 5 and beyond, let us rewrite Eqs. (4.6.1) and (4.6.2) with emphasis on the bilinearity and linearity of these functionals. For the elastic potential energy, kinetic energy and external work, we set

$$a(u, v) = \int_0^{t_T} k u v dt, \quad (4.6.9)$$

$$b(u, v) = \int_0^{t_T} m u v dt, \quad (4.6.10)$$

$$l(v) = \int_0^{t_T} p v dx - m \beta v(t_T). \quad (4.6.11)$$

Then, Eq. (4.6.1) can be written as

$$f(u) = \frac{1}{2}b(\dot{u}, \dot{u}) - \frac{1}{2}a(u, u) + l(u). \quad (4.6.12)$$

If these definitions are used, the problem of seeking the displacement of spring mass system based on the expanded Hamilton's principle can be rewritten as follows.

Problem 4.6.1 (Stationary Problem of Extended Action Integral) Let $V = \{v \in H^1((0, t_T); \mathbb{R}) \mid v(0) = 0\}$. Let a , b and l be given Eqs. (4.6.9), (4.6.10) and (4.6.11), respectively. Suppose $u_0 \in H^1((0, t_T); \mathbb{R})$ satisfies $u_0(0) = \alpha$. Find $u - u_0 \in V$ at which $f(u)$ of Eq. (4.6.12) is stationary. \square

When the time t_T is taken sufficiently smaller than half cycle of the natural vibration $\pi/\sqrt{k/m}$, the stationary condition can be changed to the minimum condition that indicates the existence of a unique minimum point [52, Section 36.2, p. 159].

Furthermore, the validity of the equation $f'(u)[v] = 0$ for an arbitrary $v \in V$ is equivalent to the equation

$$a(u, v) - b(\dot{u}, \dot{v}) = l(v) \quad (4.6.13)$$

that holds for any $v \in V$. Hence, Problem 4.6.1 can also be expressed as follows.

Problem 4.6.2 (Variational Problem of Extended Action Integral) Let $V = \{v \in H^1((0, t_T); \mathbb{R}) \mid v(0) = 0\}$. Let a , b and l be given by Eqs. (4.6.9), (4.6.10) and (4.6.11), respectively. Suppose $u_0 \in H^1((0, t_T); \mathbb{R})$ satisfies $u_0(0) = \alpha$. Find $u - u_0 \in V$ satisfying Eq. (4.6.13) with respect to an arbitrary $v \in V$. \square

4.6.2 Minimum Principle of Potential Energy

Let us consider a function space with respect to functions used in the minimum principle of potential energy (Problem 4.1.2). Potential energy was defined in Eq. (4.1.8) and can be rewritten as follows:

$$\pi(u) = \int_0^l \frac{1}{2}e_Y \nabla u \nabla u a_S dx - \int_0^l bu a_S dx - p_N u(l) a_S(l) \quad (4.6.14)$$

for arbitrary $u \in U$. Moreover, as a stationary condition of $\pi(u + v)$ with respect to an arbitrary $v \in U$,

$$\pi'(u)[v] = \int_0^l (e_Y \nabla u \nabla v - bv) a_S dx - p_N v(l) a_S(l) = 0 \quad (4.6.15)$$

was obtained during the calculation of Eq. (4.1.9). In order for these integrals to make sense, relationships such as those seen in Eqs. (4.6.5), (4.6.6) and (4.6.7) should be used and the conditions

$$u, v \in U = \left\{ u \in H^1((0, l); \mathbb{R}) \mid u(0) = 0 \right\}, \quad e_Y \in L^\infty((0, l); \mathbb{R}), \\ b \in L^2((0, l); \mathbb{R}), \quad a_S \in W^{1,\infty}((0, l); \mathbb{R})$$

should be assumed. Here, a_S needs to be $W^{1,\infty}$ class in the neighborhood of $x = l$ in order to use its boundary value.

With these assumptions, the integrals of Eqs. (4.6.14) and (4.6.15) have meanings. In this situation, $\pi'(u)[v]$ becomes a bounded linear functional with respect to $v \in U$ and can be written as

$$\pi'(u)[v] = \langle g, v \rangle_{U' \times U}, \quad (4.6.16)$$

where g is an element of U' and is called the gradient of π .

We shall again concentrate on the bilinearity or linearity of u and v which are elastic potential energy and external work and consider the functionals

$$a(u, v) = \int_0^l e_Y \nabla u \cdot \nabla v a_S \, dx, \quad (4.6.17)$$

$$l(v) = \int_0^l b v a_S \, dx + a_S(l) p_N v(l). \quad (4.6.18)$$

Hence, Eq. (4.6.14) can be written as

$$\pi(u) = \frac{1}{2} a(u, u) - l(u). \quad (4.6.19)$$

If these definitions are used, the problem for seeking the displacement of one-dimensional elastic bodies can be written as follows.

Problem 4.6.3 (Minimization Problem of Potential Energy) Let $U = \{v \in H^1((0, l); \mathbb{R}) \mid v(0) = 0\}$. Suppose π is given by Eq. (4.6.19). In this case, find an element u which satisfies

$$\min_{u \in U} \pi(u).$$

□

The fact that $\pi'(u)[v] = 0$ holds with respect to an arbitrary $v \in U$ is equivalent to

$$a(u, v) = l(v) \quad (4.6.20)$$

that holds for any $v \in U$. In this case, Problem 4.6.3 can also be rewritten in the following way.

Problem 4.6.4 (Variational Problem of Potential Energy) Let $U = \{v \in H^1((0, l); \mathbb{R}) \mid v(0) = 0\}$. Let a and l be given by Eqs. (4.6.17) and (4.6.18), respectively. In this case, find a $u \in U$ which satisfies Eq. (4.6.20) with respect to an arbitrary $v \in U$. \square

The existence of unique solutions to Problem 4.6.3 and Problem 4.6.4 is shown in Sect. 5.2.

4.6.3 Pontryagin's Minimum Principle

With respect to the optimum control problem (Problem 4.1.4) of a linear system, the Lagrange function was defined as Eq. (4.1.21) and can be rewritten as follows:

$$\begin{aligned} \mathcal{L}(\xi, u, z_0, p) &= \mathcal{L}_0(\xi, u, z_0) + \mathcal{L}_1(\xi, p) \\ &= f_0(\xi, u) - \int_0^{t_T} (\dot{u} - Au - B\xi) \cdot z \, dt + \int_0^{t_T} \left(\frac{\|\xi\|_{\mathbb{R}^d}^2}{2} - 1 \right) p \, dt \\ &= \int_0^{t_T} \left\{ \frac{\|u\|_{\mathbb{R}^n}^2}{2} + \frac{\|\xi\|_{\mathbb{R}^d}^2}{2} - (\dot{u} - Au - B\xi) \cdot z + \left(\frac{\|\xi\|_{\mathbb{R}^d}^2}{2} - 1 \right) p \right\} \, dt \\ &\quad + \frac{1}{2} \|u(t_T)\|_{\mathbb{R}^n}^2 \end{aligned} \quad (4.6.21)$$

with respect to $(\xi, u, z_0, p) \in \Xi \times U \times Z \times P$. Moreover, the first variation of \mathcal{L} can be summarized as

$$\begin{aligned} \mathcal{L}'(\xi, u, z_0, p) [\eta, \hat{u}, \hat{z}_0, \hat{p}] &= \mathcal{L}_\xi(\xi, u, z_0, p) [\eta] + \mathcal{L}_u(\xi, u, z_0, p) [\hat{u}] + \mathcal{L}_{z_0}(\xi, u, z_0, p) [\hat{z}_0] \\ &\quad + \mathcal{L}_p(\xi, u, z_0, p) [\hat{p}] \end{aligned} \quad (4.6.22)$$

with respect to an arbitrary variation $(\eta, \hat{u}, \hat{z}_0, \hat{p}) \in \Xi \times V \times W \times P$ of $(\xi, u, z_0, p) \in \Xi \times U \times Z \times P$, where

$$\mathcal{L}_\xi(\xi, u, z_0, p) [\eta] = \int_0^{t_T} \left((1 + p) \xi + B^\top z_0 \right) \cdot \eta \, dt = \langle g, \eta \rangle, \quad (4.6.23)$$

$$\begin{aligned} \mathcal{L}_u(\xi, u, z_0, p) [\hat{u}] &= \int_0^{t_T} \left(u + \dot{z}_0 + A^\top z_0 \right) \cdot \hat{u} \, dt \\ &\quad + (u(t_T) - z_0(t_T)) \cdot \hat{u}(t_T), \end{aligned} \quad (4.6.24)$$

$$\mathcal{L}_{z_0}(\xi, \mathbf{u}, z_0, p)[\hat{z}_0] = - \int_0^{t_T} (\dot{\mathbf{u}} - A\mathbf{u} - B\xi) \cdot \hat{z}_0 \, dt, \quad (4.6.25)$$

$$\mathcal{L}_p(\xi, \mathbf{u}, z_0, p)[q] = \int_0^{t_T} \left(\frac{\|\xi\|_{\mathbb{R}^d}^2}{2} - 1 \right) q \, dt. \quad (4.6.26)$$

In order for the integrals of Eqs. (4.6.21) and (4.6.22) to make sense, the assumptions

$$\begin{aligned} \Xi &= L^2((0, t_T); \mathbb{R}^d), \\ U &= \left\{ \mathbf{u} \in H^1((0, t_T); \mathbb{R}^n) \mid \mathbf{u}(0) = \boldsymbol{\alpha} \right\}, \\ V &= \left\{ \mathbf{v} \in H^1((0, t_T); \mathbb{R}^n) \mid \mathbf{v}(0) = \mathbf{0}_{\mathbb{R}^n} \right\}, \\ Z &= \left\{ \mathbf{z} \in H^1((0, t_T); \mathbb{R}^n) \mid \mathbf{z}(t_T) = \mathbf{u}(t_T) \right\}, \\ W &= \left\{ \mathbf{w} \in H^1((0, t_T); \mathbb{R}^n) \mid \mathbf{w}(t_T) = \mathbf{0}_{\mathbb{R}^n} \right\} \end{aligned}$$

need to be made based on relationships such as those seen in Sect. 4.6.1. Here, U and Z are affine subspaces with respect to V and W respectively. If elements \mathbf{u}_0 and \mathbf{z}_T of $H^1((0, t_T); \mathbb{R})$ which satisfy $\mathbf{u}(0) = \boldsymbol{\alpha}$ and $\mathbf{z}(t_T) = \mathbf{u}(t_T)$, respectively, are selected and fixed, U and Z are equivalent to $V(\mathbf{u}_0)$ and $W(\mathbf{z}_T)$, respectively.

If definitions such as those above are used, Problem 4.1.3 can be written as follows.

Problem 4.6.5 (Linear System with Control) Let $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times d}$, $\boldsymbol{\alpha} \in \mathbb{R}^n$ and control force $\xi \in \Xi$ be given. Find $\mathbf{u} - \mathbf{u}_0 \in V$ such that Eq. (4.6.25) is zero with respect to an arbitrary $\hat{z}_0 \in W$. \square

The adjoint problem determining z_0 (Problem 4.1.5) can be written in the following way.

Problem 4.6.6 (Adjoint Problem with Respect to f_0) Let $A \in \mathbb{R}^{n \times n}$ be given as in Problem 4.6.5. Find $z_0 - z_T \in W$ such that Eq. (4.6.24) is zero with respect to an arbitrary $\hat{\mathbf{u}} \in V$. \square

A similar expression is possible with respect to an optimum control problem of non-linear systems (Problem 4.1.7) but it is omitted here.

4.7 Summary

In Chap. 4, we looked at what an optimization problem is when a design variable is a function defined on the domain expressing time or location (function optimization problem) by relating to the basics of functional analysis. The key points in this chapter are listed as follows:

- (1) The equation of motion of a one-degree-of-freedom spring mass system can be obtained as the stationary condition (Hamilton's principle) of an action integral (Sect. 4.1.1). Moreover, the elasticity equation of a one-dimensional elastic body can be obtained as the minimum condition (minimum principle of potential energy) of potential energy (Sect. 4.1.2). Furthermore, the optimum solution for optimum control problem can be given as the minimal condition (Pontryagin's minimizing principle) of the Hamiltonian (Sect. 4.1.3).
- (2) A linear space (vector space) is a set such that all the linear combinations among all elements are included. The set of all continuous functions is a linear space (Sect. 4.2.1).
- (3) The linear space for which a norm is defined is called the norm space. Furthermore, a linear space which is complete (all Cauchy sequences converge) with respect to the norm is called a Banach space. The set of all continuous functions becomes a Banach space with the maximum value as the norm (Sect. 4.2.4).
- (4) A linear space in which an inner product is defined is called the inner product space. Furthermore, a linear space which is complete with the norm defined by an inner product is called a Hilbert space. A finite-dimensional vector space is a Hilbert space (Sect. 4.2.5).
- (5) Function spaces of the Hölder space $C^{k,\sigma}(\Omega; \mathbb{R})$, Lebesgue space $L^p(\Omega; \mathbb{R})$ and Sobolev space $W^{k,p}(\Omega; \mathbb{R})$ are Banach spaces with their corresponding norms. Moreover, $L^2(\Omega; \mathbb{R})$ and $H^k(\Omega; \mathbb{R}) = W^{k,2}(\Omega; \mathbb{R})$ are Hilbert spaces (Sect. 4.3). The embedding relationships of these function spaces are given by Sobolev embedding theorems (Theorem 4.3.14).
- (6) The set of all bounded linear functionals on a Banach space is called a dual space. The Fréchet derivative of a functional is defined by a dual product of variation vector and gradient (Sect. 4.4.6).
- (7) Hamilton's principle, minimum principle of potential energy and optimum control problems can be defined as function optimization problems on function spaces $H^1((0, t_T); \mathbb{R})$, $H^1((0, l); \mathbb{R})$ and $L^2((0, t_T); \mathbb{R}^d)$ respectively (Sect. 4.6).

4.8 Practice Problems

- 4.1** Introducing time $t \in (0, t_T)$ into Problem 4.1.2 representing the minimum principle of potential energy with respect to a one-dimensional elastic body, show

the equation of motion and the terminal condition of the velocity via the expanded Hamilton's principle using the following order. In this case, let $\rho : (0, l) \rightarrow \mathbb{R}$ ($\rho > 0$) be the density, $\alpha : (0, l) \rightarrow \mathbb{R}$ be displacement when $t = 0$, $\beta : (0, l) \rightarrow \mathbb{R}$ be the velocity at $t = t_T$, $b : (0, l) \times (0, t_T) \rightarrow \mathbb{R}$ be the volume force and $p_N : (0, t_T) \rightarrow \mathbb{R}$ be the boundary force.

- Define U as a set of displacements $u : (0, l) \times (0, t_T) \rightarrow \mathbb{R}$ that satisfy

$$u(0, t) = 0 \quad t \in (0, t_T), \quad u(x, 0) = \alpha(x) \quad x \in (0, l).$$

Define V as a set of variational displacement v expressing an arbitrary variation of $u \in U$.

- Let

$$\begin{aligned} f(u) = \int_0^{t_T} \left\{ \int_0^l \left(\frac{1}{2} \rho \dot{u}^2 - \frac{1}{2} e_Y (\nabla u)^2 + bu \right) a_S dx \right. \\ \left. + p_N u(l, t) a_S(l) \right\} dt - \int_0^l \rho \beta u(x, t_T) a_S dx \end{aligned}$$

be an extended action integral and seek the stationary condition of f for an arbitrary $v \in V$.

- Determine the appropriate function spaces for ρ, α, β, b and p_N .

4.2 Function spaces relating to a generalized displacement \mathbf{u} of an $n \in \mathbb{N}$ -degree-of-freedom system and its variation \mathbf{v} are set to be

$$\begin{aligned} U &= \left\{ \mathbf{u} \in H^1((0, t_T); \mathbb{R}^n) \mid \mathbf{u}(0) = \boldsymbol{\alpha}, \mathbf{u}(t_T) = \boldsymbol{\beta} \right\}, \\ V &= \left\{ \mathbf{v} \in H^1((0, t_T); \mathbb{R}^n) \mid \mathbf{v}(0) = \mathbf{0}_{\mathbb{R}^n}, \mathbf{v}(t_T) = \mathbf{0}_{\mathbb{R}^n} \right\}, \end{aligned}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are elements of \mathbb{R}^n . Suppose the kinetic energy $\kappa(\mathbf{u}, \dot{\mathbf{u}})$ and potential energy $\pi(\mathbf{u}, \dot{\mathbf{u}})$ with respect to $\mathbf{u} \in U$ are given. Moreover, suppose that the Lagrange function in mechanics be defined by $l(\mathbf{u}, \dot{\mathbf{u}}) = \kappa(\mathbf{u}, \dot{\mathbf{u}}) - \pi(\mathbf{u}, \dot{\mathbf{u}})$. Furthermore, assume that the action integral is defined by

$$f(\mathbf{u}, \dot{\mathbf{u}}) = \int_0^{t_T} l(\mathbf{u}, \dot{\mathbf{u}}) dt.$$

Show that the Lagrange equation of motion

$$\frac{d}{dt} \frac{\partial l}{\partial \dot{\mathbf{u}}} - \frac{\partial l}{\partial \mathbf{u}} = \mathbf{0}_{\mathbb{R}^n}$$

can be obtained from the stationary condition (Hamilton's principle) of $f(\mathbf{u} + \mathbf{v}, \dot{\mathbf{u}} + \dot{\mathbf{v}})$ with respect to an arbitrary $\mathbf{v} \in V$.

4.3 Introducing a generalized momentum $\mathbf{q} \in Q = H^1((0, t_T); \mathbb{R}^n)$ to Practice 4.2 and calling $\mathcal{H}(\mathbf{u}, \mathbf{q}) = -l(\mathbf{u}, \mathbf{q}) + \mathbf{q} \cdot \dot{\mathbf{u}}$ the Hamiltonian, let

$$f(\mathbf{u}, \mathbf{q}) = \int_0^{t_T} (-\dot{\mathbf{q}} \cdot \mathbf{u} - \mathcal{H}(\mathbf{u}, \mathbf{q})) dt$$

be the action integral. In this case, show that the stationary condition of $f(\mathbf{u}, \mathbf{q})$ becomes Hamilton's equation of motion

$$\dot{\mathbf{q}} = -\frac{\partial \mathcal{H}}{\partial \mathbf{u}}, \quad \dot{\mathbf{u}} = \frac{\partial \mathcal{H}}{\partial \mathbf{q}}.$$

Moreover, when Hamilton's equation of motion holds, show that $\dot{\mathcal{H}}(\mathbf{u}, \mathbf{q}) = 0$ (the Hamiltonian is conserved). Furthermore, find $\mathcal{H}(\mathbf{u}, \mathbf{q})$ with respect to the spring mass system of Fig. 4.1 assuming an external force $p = 0$.

4.4 Let Y and Z be Banach spaces and Y be compactly embedded within Z ($Y \Subset Z$) (Theorem 4.4.15). Show that $Z' \Subset Y'$ holds with respect to the dual spaces Y' and Z' of Y and Z .

Chapter 5

Boundary Value Problems of Partial Differential Equations



As seen in Chap. 1, optimal design problems are optimization problems whose state equations are considered as equality constraints. In Chap. 1, we have considered design variables and state variables as elements of a finite-dimensional vector space. However, in this book, our main interest focuses on the shape optimization problem of continuum. In this case, boundary value problems of partial differential equations, such as linear elastic bodies and Stokes flow field, are included in the equality constraints as state equations.

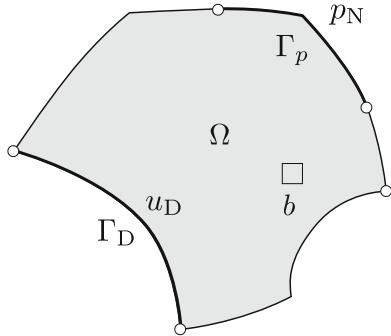
In this chapter, the definitions and the results of functional analyses discussed in Chap. 4 are used to study the methods of expressing boundary value problems of elliptic partial differential equation in their corresponding variational form (here on referred to as the weak form) as well as theorems relating to the existence of unique solutions. This weak form is not only used when considering methods in numerical analysis with respect to boundary value problems of elliptic partial differential equations shown in Chap. 6, but also as Lagrange functions with respect to shape and topology optimization problems where boundary value problems are included in the equality constraint (see Chaps. 8 and 9).

5.1 Poisson Problem

Let us consider a Poisson problem as a simple example of a boundary value problem of an elliptic partial differential equation (Definition A.7.1) and look at its definition and the process of transforming the system in its weak form. A Poisson problem can be thought of, for instance, as a situation when thermal conductivity is 1 in a stationary heat conduction problem (Sect. A.6).

Let the domain Ω be a Lipschitz domain (Sect. A.5) of $d \in \{2, 3\}$ dimensions such as in Fig. 5.1. Let Γ_D be a partial open set of $\partial\Omega$, the boundary of Ω at which temperature is given in the heat conductivity problem. The remaining boundary

Fig. 5.1 Domain Ω and its boundary $\partial\Omega = \bar{\Gamma}_D \cup \bar{\Gamma}_N$



$\Gamma_N = \partial\Omega \setminus \bar{\Gamma}_D$ is taken to be the boundary at which heat flux is given. Furthermore, let $\Gamma_p \subset \Gamma_N$ be the boundary at which the heat flow is non-zero. In this chapter Γ_p and $\Gamma_N \setminus \Gamma_p$ are not distinguished from one another but they will be considered separately in Chap. 9. Moreover, $\Delta = \nabla \cdot \nabla$ expresses the Laplace operator. Meanwhile, ν expresses the outward unit normal vector defined on the boundary (Definition A.5.4) and $\partial_\nu = \nu \cdot \nabla$. In this case, a Poisson problem with mixed boundary conditions is defined as follows.

Problem 5.1.1 (Poisson Problem) Let the functions $b : \Omega \rightarrow \mathbb{R}$, $p_N : \Gamma_N \rightarrow \mathbb{R}$, and $u_D : \Omega \rightarrow \mathbb{R}$ be given. Find the function $u : \Omega \rightarrow \mathbb{R}$ such that the system

$$-\Delta u = b \quad \text{in } \Omega, \quad (5.1.1)$$

$$\partial_\nu u = p_N \quad \text{on } \Gamma_N, \quad (5.1.2)$$

$$u = u_D \quad \text{on } \Gamma_D, \quad (5.1.3)$$

is satisfied. \square

In Problem 5.1.1, Eq. (5.1.1) is called a Poisson equation. Moreover, when $b = 0$, it is called a Laplace equation or homogeneous Poisson equation. The problem in that case (Problem 5.1.1) is called a Laplace problem.

Moreover, the boundary condition of Eq. (5.1.3) expresses the relationship established by the traces on Γ_D of the function u and u_D defined on Ω . In this situation, there is a need to define an appropriate function space of u and u_D such that a trace can be taken. On the other hand, $\partial_\nu u$ of Eq. (5.1.2) shows the relationship of a trace on Γ_N . In order for this relationship to have a meaning, assumptions such that trace on the boundary of ∇u can be taken have to be specified. However, as shown below, if Problem 5.1.1 is changed to an integral equation (weak form), it should be noted that such an assumption becomes unnecessary.

From the above considerations, u_D is assumed to be an element of $H^1(\Omega; \mathbb{R})$ and the set of functions satisfying Eq. (5.1.3) is taken to be

$$U(u_D) = \left\{ v \in H^1(\Omega; \mathbb{R}) \mid v = u_D \text{ on } \Gamma_D \right\}.$$

As seen in Sect. 4.6, $U (u_D)$ is an affine subspace of the Hilbert space

$$U = \left\{ v \in H^1(\Omega; \mathbb{R}) \mid v = 0 \text{ on } \Gamma_D \right\}. \quad (5.1.4)$$

The fact that U is a Hilbert space will be needed when fitting the Poisson problem into the framework of abstract variational problem later on.

If both sides of Eq. (5.1.1) are multiplied by an arbitrary $v \in U$ and integrated over Ω , the Gauss–Green theorem (Theorem A.8.2) can be employed to establish

$$-\int_{\Omega} \Delta u v \, dx = \int_{\Omega} \nabla u \cdot \nabla v \, dx - \int_{\Gamma_N} \partial_{\nu} u v \, d\gamma = \int_{\Omega} b v \, dx, \quad (5.1.5)$$

where the fact that $v = 0$ on Γ_D was used. On the other hand, if an arbitrary $v \in U$ is multiplied to both sides of Eq. (5.1.2) and integrated over Γ_N , then the equation

$$\int_{\Gamma_N} \partial_{\nu} u v \, d\gamma = \int_{\Gamma_N} p_N v \, d\gamma \quad (5.1.6)$$

is established. Here, substituting Eq. (5.1.6) into Eq. (5.1.5) gives

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} b v \, dx + \int_{\Gamma_N} p_N v \, d\gamma. \quad (5.1.7)$$

Equation (5.1.7) is called a weak form of a Poisson problem.

Let us mention in advance the fact that the arbitrary function $v \in U$ used when obtaining the weak-form equation will be used as a Lagrange multiplier with respect to a boundary value problem when considering a shape or topology optimization problem in which a boundary value problem is included in the equality constraints.

Furthermore, the left-hand side of Eq. (5.1.7) has the property of being bilinear with respect to u and v . Moreover, the right-hand side of Eq. (5.1.7) is linear with respect to v . Here, as was seen in Sect. 4.6 we define

$$a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v \, dx, \quad (5.1.8)$$

$$l(v) = \int_{\Omega} b v \, dx + \int_{\Gamma_N} p_N v \, d\gamma. \quad (5.1.9)$$

Using these definitions, the weak form of Problem 5.1.1 is given as follows.

Problem 5.1.2 (Weak Form of Poisson Problem) Let U be defined as in Eq. (5.1.4) and the consider the functions $b \in L^2(\Omega; \mathbb{R})$, $p_N \in L^2(\Gamma_N; \mathbb{R})$ and $u_D \in H^1(\Omega; \mathbb{R})$. Moreover, let $a(\cdot, \cdot)$ and $l(\cdot)$ be given by Eqs. (5.1.8) and (5.1.9), respectively. Find u such that $\tilde{u} = u - u_D \in U$ satisfying

$$a(u, v) = l(v), \quad (5.1.10)$$

for all $v \in U$. □

Here, let us compare Problem 5.1.1 with Problem 5.1.2. In Problem 5.1.1, in order for Eq. (5.1.1) to have meaning, u needs to be second-order differentiable. Moreover, $\partial_v u$ needs to be defined on Γ_N . On the other hand, in Problem 5.1.2, there is no need for u to be second-order differentiable, instead in order for the integral of Eq. (5.1.8) to be defined, there is a need for the first-order derivative of both u and v to be square integrable. In this way, depending on the conditions that the solutions should satisfy, Problem 5.1.1 is referred to as the strong form of Poisson problem and Problem 5.1.2 as the weak form of Poisson problem. Moreover, the solution u of Problem 5.1.2 is called the weak solution. Furthermore, as shown in Sect. 5.2, the fact that a unique solution exists is guaranteed by the weak solution.

The following terminology is used in a boundary value problem of a differential equation:

- Equation (5.1.3) is called a Dirichlet condition or fundamental boundary condition or first-type boundary condition. The boundary for which the Dirichlet condition is given is called a Dirichlet boundary. Problem 5.1.1 or Problem 5.1.2 for which this condition is given over the entire boundary is called a Dirichlet problem.
- Equation (5.1.2) can also be called a Neumann condition or natural boundary condition or second-type boundary condition. The boundary with Neumann condition is called a Neumann boundary. Problem 5.1.1 or Problem 5.1.2 with this condition given over the entire boundary is called a Neumann problem. However, there is a need to note that a Neumann problem does not have a unique solution (Exercise 5.2.6).
- When both the Dirichlet condition and Neumann condition exist, it is referred to as mixed boundary value problem.
- For the Dirichlet condition or Neumann condition, if $u_D = 0$ or $p_N = 0$ respectively, it is called a homogeneous type. When $u_D \neq 0$ or $p_N \neq 0$, it is called an inhomogeneous type.

5.1.1 *Extended Poisson Problem*

Let us consider an extended Poisson problem. This problem is used when specifying an abstract gradient method in Chap. 8. Moreover, a linear elastic problem extended in a similar manner to that shown here will be used when specifying an abstract gradient method in Chap. 9 too.

We use the symbols used in Problem 5.1.1 to extend the Poisson problem in the following way.

Problem 5.1.3 (Extended Poisson Problem) Let the functions $b : \Omega \rightarrow \mathbb{R}$, $c_\Omega : \Omega \rightarrow \mathbb{R}$, $p_R : \partial\Omega \rightarrow \mathbb{R}$ and $c_{\partial\Omega} : \partial\Omega \rightarrow \mathbb{R}$ be given. Find the function $u : \Omega \rightarrow \mathbb{R}$ satisfying

$$-\Delta u + c_\Omega u = b \quad \text{in } \Omega, \quad (5.1.11)$$

$$\partial_\nu u + c_{\partial\Omega} u = p_R \quad \text{on } \partial\Omega. \quad (5.1.12)$$

□

In Problem 5.1.3, Eq. (5.1.12) is called a Robin condition or third-type boundary condition. In reference to Problem 5.1.3, the problem when this condition is given across the entire boundary is called a Robin problem.

The weak form of Problem 5.1.3 can be obtained in the following way. Here

$$U = H^1(\Omega; \mathbb{R}) \quad (5.1.13)$$

is set. Multiplying both sides of Eq. (5.1.11) by an arbitrary $v \in U$ and integrating over Ω , then using the Gauss–Green theorem (Theorem A.8.2) gives

$$\begin{aligned} \int_{\Omega} (-\Delta u + c_\Omega u) v \, dx &= \int_{\Omega} (\nabla u \cdot \nabla v + c_\Omega u v) \, dx - \int_{\partial\Omega} \partial_\nu u v \, d\gamma \\ &= \int_{\Omega} b v \, dx. \end{aligned} \quad (5.1.14)$$

On the other hand, if both sides of Eq. (5.1.12) are multiplied by an arbitrary $v \in U$ and integrated over $\partial\Omega$, the equality

$$\int_{\partial\Omega} \partial_\nu u v \, d\gamma = \int_{\partial\Omega} (p_R - c_{\partial\Omega} u) v \, d\gamma \quad (5.1.15)$$

holds. Here, if Eq. (5.1.15) is substituted into Eq. (5.1.14), the equation

$$\int_{\Omega} (\nabla u \cdot \nabla v + c_\Omega u v) \, dx + \int_{\partial\Omega} c_{\partial\Omega} u v \, d\gamma = \int_{\Omega} b v \, dx + \int_{\partial\Omega} p_R v \, d\gamma \quad (5.1.16)$$

is obtained. Eq. (5.1.16) is the weak form of Problem 5.1.3.

Here, note that the left-hand side of Eq. (5.1.16) is bilinear with respect to u and v and the right-hand side is linear with respect to v . Let $a : U \times U \rightarrow \mathbb{R}$ and $l : U \rightarrow \mathbb{R}$ be

$$a(u, v) = \int_{\Omega} (\nabla u \cdot \nabla v + c_\Omega u v) \, dx + \int_{\partial\Omega} c_{\partial\Omega} u v \, d\gamma, \quad (5.1.17)$$

$$l(v) = \int_{\Omega} b v \, dx + \int_{\partial\Omega} p_R v \, d\gamma. \quad (5.1.18)$$

Here, the weak form of Problem 5.1.3 becomes as follows.

Problem 5.1.4 (Weak Form of Extended Poisson Problem) Let U be Eq. (5.1.13) and the functions $b \in L^2(\Omega; \mathbb{R})$, $c_\Omega \in L^\infty(\Omega; \mathbb{R})$, $p_R \in L^2(\partial\Omega; \mathbb{R})$, $c_{\partial\Omega} \in L^\infty(\partial\Omega; \mathbb{R})$. Moreover, let $a(\cdot, \cdot)$ and $l(\cdot)$ be given by Eqs. (5.1.17) and (5.1.18), respectively. In this case, obtain a $u \in U$ which satisfies

$$a(u, v) = l(v) \quad (5.1.19)$$

with respect to an arbitrary $v \in U$. \square

5.2 Abstract Variational Problem

In Sects. 5.1 and 5.1.1, the weak forms of the Poisson problem and extended Poisson problem were shown as Eqs. (5.1.10) and (5.1.19), respectively. These are classified as boundary value problems of elliptic partial differential equations based on classification of linear second-order partial differential equations (Definition A.7.1). If it is a boundary value problem of an elliptic partial differential equation, it can be expected that either of the weak forms can be expressed using a bilinear form a and linear form l . Hence, let us define an abstract variational problem which abstracts the weak form of elliptic partial differential equation and investigate the existence of a unique solution to such a problem.

In this section, U is taken to be a real Hilbert space. Let us define two characteristics with respect to a bilinear form on U (Sect. 4.4).

Definition 5.2.1 (Coercive Bilinear Form on Real Hilbert Space) Let $a : U \times U \rightarrow \mathbb{R}$ be a bilinear form on U . If some constant $\alpha > 0$ exists with respect to an arbitrary $v \in U$ and

$$a(v, v) \geq \alpha \|v\|_U^2$$

holds, a is referred to as coercive or elliptic. \square

If U is \mathbb{R}^d , the bilinear form equation with respect to $x, y \in \mathbb{R}^d$ can be written as $a(x, y) = x \cdot (Ay)$. Here, A is a matrix of $\mathbb{R}^{d \times d}$. When $A = A^\top$, coerciveness of a is equivalent to A being positive definite.

Definition 5.2.2 (Boundedness of Bilinear Form on Real Hilbert Space) Let $a : U \times U \rightarrow \mathbb{R}$ be a bilinear form on U . If there exists some $\beta > 0$ with respect to an arbitrary $u, v \in U$ and

$$|a(u, v)| \leq \beta \|u\|_U \|v\|_U$$

holds, a is said to be bounded. \square

If $U = \mathbb{R}^d$, the boundedness of the bilinear form $a(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot (\mathbf{A}\mathbf{y})$ becomes equivalent to the norm of matrix \mathbf{A} being bounded (see Eq. (4.4.3)).

Let us consider the next problem using the definitions above.

Problem 5.2.3 (Abstract Variational Problem) Let $a : U \times U \rightarrow \mathbb{R}$ be a bilinear form on U and $l = l(\cdot) = \langle l, \cdot \rangle \in U'$. In this case, obtain a $u \in U$ which satisfies

$$a(u, v) = l(v)$$

with respect to an arbitrary $v \in U$. \square

Let $U = \mathbb{R}^d$. If the matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ in the bilinear form $a(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot (\mathbf{A}\mathbf{y})$ and $\mathbf{b} \in \mathbb{R}^d$ are given, an abstract variational problem becomes a problem seeking $\mathbf{x} \in \mathbb{R}^d$ which satisfies

$$(\mathbf{A}\mathbf{x}) \cdot \mathbf{y} = \mathbf{b} \cdot \mathbf{y} \quad (5.2.1)$$

with respect to an arbitrary $\mathbf{y} \in \mathbb{R}^d$.

5.2.1 Lax–Milgram Theorem

The fact that there exists a unique solution to Problem 5.2.3 is guaranteed by the Lax–Milgram theorem. In this theorem, it is assumed that a bilinear form a is coercive and bounded. Since these characteristics are the same as the definition of an inner product in Hilbert spaces, this theorem is proven using Riesz’s representation theorem (Theorem 4.4.17) (cf. [42, Theorem 1.3, p. 29], [48, Theorem 1, p. 297], [158, Theorem 2.6, p. 48]).

Theorem 5.2.4 (Lax–Milgram Theorem) *In Problem 5.2.3, let a be coercive and bounded. Moreover, let $l \in U'$. In this case, there is a unique solution $u \in U$ for Problem 5.2.3 and*

$$\|u\|_U \leq \frac{1}{\alpha} \|l\|_{U'}$$

holds with respect to α used in Definition 5.2.1. \square

If $U = \mathbb{R}^d$, assuming \mathbf{A} is symmetric bounded and positive definite, there exists an inverse matrix to \mathbf{A} and \mathbf{x} satisfying Eq. (5.2.1) becomes

$$\mathbf{x} = \mathbf{A}^{-1} \mathbf{b}. \quad (5.2.2)$$

Here, the inequality

$$\|\mathbf{x}\|_{\mathbb{R}^d} \leq \frac{1}{\alpha} \|\mathbf{b}\|_{\mathbb{R}^d}$$

holds, where α is the minimum eigenvalue of \mathbf{A} . Moreover, when \mathbf{A} is asymmetric, the positive definiteness of \mathbf{A} is replaced with that of $(\mathbf{A}^\top + \mathbf{A})/2$, because $\mathbf{x} \cdot \{(\mathbf{A}^\top + \mathbf{A})\mathbf{x}\} \geq 2\alpha \|\mathbf{x}\|_{\mathbb{R}^d}^2$ holds if $\mathbf{x} \cdot (\mathbf{A}\mathbf{x}) \geq \alpha \|\mathbf{x}\|_{\mathbb{R}^d}^2$ with respect to an arbitrary $\mathbf{x} \in \mathbb{R}^d$.

Next, let us show the existence of a unique solution to the Poisson problem using the Lax–Milgram theorem.

Exercise 5.2.5 (Existence of Unique Solution to Poisson Problem) In Problem 5.1.2, when $|\Gamma_D| (= \int_{\Gamma_D} d\gamma)$ is positive, show that there exists a unique solution $\tilde{u} = u - u_D \in U$. \square

Answer The assumptions of the Lax–Milgram theorem with respect to Problem 5.1.2 need to be shown. Consider the Hilbert space $U = \{u \in H^1(\Omega; \mathbb{R}) \mid u = 0 \text{ on } \Gamma_D\}$. Moreover, if we let

$$\hat{l}(v) = l(v) - a(u_D, v), \quad (5.2.3)$$

Problem 5.1.2 can be written as a problem seeking $\tilde{u} = u - u_D \in U$ which satisfies

$$a(\tilde{u}, v) = \hat{l}(v),$$

with respect to an arbitrary $v \in U$. In view of these relationships, the assumptions of the Lax–Milgram theorem hold in the following ways:

(1) a is coercive. In fact,

$$a(v, v) = \int_{\Omega} \nabla v \cdot \nabla v \, dx = \|\nabla v\|_{L^2(\Omega; \mathbb{R}^d)}^2 \geq \frac{1}{c^2} \|v\|_{H^1(\Omega; \mathbb{R})}^2$$

holds because of Poincaré’s inequality (Corollary A.9.4). If we let $1/c^2$ be α , from Definition 5.2.1, a is coercive.

(2) a is bounded. In fact, using Hölder’s inequality (Theorem A.9.1), the inequality

$$\begin{aligned} |a(u, v)| &= \left| \int_{\Omega} \nabla u \cdot \nabla v \, dx \right| \leq \|\nabla u\|_{L^2(\Omega; \mathbb{R}^d)} \|\nabla v\|_{L^2(\Omega; \mathbb{R}^d)} \\ &\leq \|u\|_{H^1(\Omega; \mathbb{R})} \|v\|_{H^1(\Omega; \mathbb{R})} \end{aligned}$$

is established. This relationship shows that it holds when $\beta = 1$ in Definition 5.2.2.

- (3) $\hat{l} \in U'$. In fact, from the fact that $\partial\Omega$ assumes a Lipschitz boundary, the norm of the trace operator (Theorem 4.4.2)

$$\|\gamma\|_{\mathcal{L}(H^1(\Omega; \mathbb{R}); H^{1/2}(\partial\Omega; \mathbb{R}))} = \sup_{v \in H^1(\Omega; \mathbb{R}) \setminus \{0_{H^1(\Omega; \mathbb{R})}\}} \frac{\|v\|_{H^{1/2}(\partial\Omega; \mathbb{R})}}{\|v\|_{H^1(\Omega; \mathbb{R})}}. \quad (5.2.4)$$

is bounded. This is set as $c_1 > 0$. Moreover, the inequalities

$$\begin{aligned} |\hat{l}(v)| &\leq \int_{\Omega} |bv| \, dx + \int_{\Gamma_N} |p_N v| \, d\gamma + \int_{\Omega} |\nabla u_D \cdot \nabla v| \, dx \\ &\leq \|b\|_{L^2(\Omega; \mathbb{R})} \|v\|_{L^2(\Omega; \mathbb{R})} + \|p_N\|_{L^2(\Gamma_N; \mathbb{R})} \|v\|_{L^2(\Gamma_N; \mathbb{R})} \\ &\quad + \|\nabla u_D\|_{L^2(\Omega; \mathbb{R}^d)} \|\nabla v\|_{L^2(\Omega; \mathbb{R}^d)} \\ &\leq (\|b\|_{L^2(\Omega; \mathbb{R})} + c_1 \|p_N\|_{L^2(\Gamma_N; \mathbb{R})} + \|u_D\|_{H^1(\Omega; \mathbb{R})}) \|v\|_{H^1(\Omega; \mathbb{R})} \end{aligned}$$

are established if Hölder's inequality is used. In Problem 5.1.2, $b \in L^2(\Omega; \mathbb{R})$, $p_N \in L^2(\Gamma_N; \mathbb{R})$ and $u_D \in H^1(\Omega; \mathbb{R})$ were assumed. Thus, the right-hand side of (\cdot) is bounded and l is a bounded linear functional on U .

Therefore, $\tilde{u} = u - u_D \in U$ exists uniquely. \square

Moreover, if the Lax–Milgram theorem is applied with respect to the Neumann problem, we get the following.

Exercise 5.2.6 (Indeterminateness of Solution to Neumann Problem) If $|\Gamma_D| = 0$ in Problem 5.1.2, show that there does not exist a unique $u \in U$ which satisfies Problem 5.1.2. Moreover, show how the problem needs to be amended in order to guarantee the existence of a unique solution. \square

Answer In the solution to Exercise 5.2.5, to show the coercivity of a , a corollary of Poincaré's inequality (Corollary A.9.4) was used from the fact that $|\Gamma_D| > 0$ was assumed. However, in Neumann problems $|\Gamma_D| = 0$, hence the corollary of Poincaré's inequality cannot be used and so a cannot be said to be coercive. Hence, the Lax–Milgram theorem cannot be used. However, if we let

$$u_D = \frac{1}{|\Omega|} \int_{\Omega} u \, dx \quad (5.2.5)$$

and use Poincaré's inequality (Theorem A.9.3), the inequality

$$a(v, v) = \int_{\Omega} \nabla v \cdot \nabla v \, dx = \|\nabla v\|_{L^2(\Omega; \mathbb{R}^d)}^2 \geq \frac{1}{c^2} \|v - u_D\|_{L^2(\Omega; \mathbb{R}^d)}^2$$

holds and a becomes coercive. Therefore, there is the existence of a unique solution if the Neumann problem is rewritten as a problem seeking $\tilde{u} = u - u_D \in U$ with respect to u_D satisfying Eq. (5.2.5). \square

From the result of Exercise 5.2.6, the solution to the Neumann problem is said to have uncertainty of constant.

Furthermore, the following assumptions are needed in order to guarantee the existence of a unique solution with respect to an extended Poisson problem (Problem 5.1.3).

Exercise 5.2.7 (Existence of Solution to Extended Poisson Problem) In Problem 5.1.4, one of the following is assumed to hold:

- (1) $c_\Omega \in L^\infty(\Omega; \mathbb{R})$ takes a positive value on most of Ω .
- (2) $c_{\partial\Omega} \in L^\infty(\partial\Omega; \mathbb{R})$ takes a positive value over most of $\partial\Omega$.

In this case, show that there exists a unique solution $u \in U$ of Problem 5.1.4. \square

Answer The assumptions of the Lax–Milgram theorem with respect to Problem 5.1.4 need to be verified. Let $U = H^1(\Omega; \mathbb{R})$ be a Hilbert space. Furthermore, the following holds:

- (1) a is coercive. In fact, from the assumption, $\text{ess inf}_{x \in \Omega} c_\Omega(x)$ and $\text{ess inf}_{x \in \partial\Omega} c_{\partial\Omega}(x)$ are set as $c_1 > 0$ and $c_2 > 0$, respectively and the norm of the inverse operator of the trace operator $\gamma : H^1(\Omega; \mathbb{R}) \rightarrow L^2(\partial\Omega; \mathbb{R})$:

$$\|\gamma^{-1}\|_{\mathcal{L}(L^2(\partial\Omega; \mathbb{R}); H^1(\Omega; \mathbb{R}))} = \sup_{v \in L^2(\partial\Omega; \mathbb{R}) \setminus \{0_{L^2(\partial\Omega; \mathbb{R})}\}} \frac{\|v\|_{H^1(\Omega; \mathbb{R})}}{\|v\|_{L^2(\partial\Omega; \mathbb{R})}}. \quad (5.2.6)$$

is set to be $c_3 > 0$,

$$\begin{aligned} a(v, v) &\geq \|\nabla v\|_{L^2(\Omega; \mathbb{R}^d)}^2 + c_1 \|v\|_{L^2(\Omega; \mathbb{R})}^2 + c_2 \|v\|_{L^2(\partial\Omega; \mathbb{R})}^2 \\ &\geq \left(\min\{1, c_1\} + \frac{c_2}{c_3^2} \right) \|v\|_{H^1(\Omega; \mathbb{R})}^2 \end{aligned}$$

holds. If (\cdot) of the right-hand side is set to be α , a becomes coercive from Definition 5.2.1.

- (2) a is bounded. In fact, when the norm $\|\gamma\|_{\mathcal{L}(H^1(\Omega; \mathbb{R}); H^{1/2}(\partial\Omega; \mathbb{R}))}$ of the trace operator of Eq. (5.2.4) is set to be c_4 ,

$$\begin{aligned} |a(u, v)| &\leq \|\nabla u\|_{L^2(\Omega; \mathbb{R}^d)} \|\nabla v\|_{L^2(\Omega; \mathbb{R}^d)} \\ &\quad + \|c_\Omega\|_{L^\infty(\Omega; \mathbb{R})} \|u\|_{L^2(\Omega; \mathbb{R})} \|v\|_{L^2(\Omega; \mathbb{R})} \\ &\quad + \|c_{\partial\Omega}\|_{L^\infty(\partial\Omega; \mathbb{R})} \|u\|_{L^2(\partial\Omega; \mathbb{R})} \|v\|_{L^2(\partial\Omega; \mathbb{R})} \\ &\leq \left(1 + \|c_\Omega\|_{L^\infty(\Omega; \mathbb{R})} + c_4^2 \|c_{\partial\Omega}\|_{L^\infty(\partial\Omega; \mathbb{R})} \right) \|u\|_{H^1(\Omega; \mathbb{R}^d)} \|v\|_{H^1(\Omega; \mathbb{R}^d)} \end{aligned}$$

is established from $c_\Omega \in L^\infty(\Omega; \mathbb{R})$ and $c_{\partial\Omega} \in L^\infty(\partial\Omega; \mathbb{R})$. If (\cdot) of the right-hand side is set to be β , a becomes bounded from Definition 5.2.2.

- (3) $l \in U'$. In fact, when the norm $\|\gamma\|_{\mathcal{L}(H^1(\Omega; \mathbb{R}); H^{1/2}(\partial\Omega; \mathbb{R}))}$ of the trace operator of Eq. (5.2.4) is set to be c_4 , the inequality

$$\begin{aligned} |l(v)| &\leq \int_{\Omega} |bv| \, dx + \int_{\partial\Omega} |p_R v| \, d\gamma \\ &\leq \|b\|_{L^2(\Omega; \mathbb{R})} \|v\|_{L^2(\Omega; \mathbb{R})} + \|p_R\|_{L^2(\partial\Omega; \mathbb{R})} \|v\|_{L^2(\partial\Omega; \mathbb{R})} \\ &\leq (\|b\|_{L^2(\Omega; \mathbb{R})} + c_4 \|p_R\|_{L^2(\partial\Omega; \mathbb{R})}) \|v\|_{H^1(\Omega; \mathbb{R})} \end{aligned}$$

is established. In Problem 5.1.4, since $b \in L^2(\Omega; \mathbb{R})$ and $p_R \in L^2(\partial\Omega; \mathbb{R})$ in (\cdot) of the right-hand side become bounded, then l becomes a bound linear functional on U .

Therefore, from the Lax–Milgram theorem, $u \in U$ exists uniquely. \square

5.2.2 Abstract Minimization Problem

In an abstract variational problem (Problem 5.2.3), if $a : U \times U \rightarrow \mathbb{R}$ is symmetric, the abstract variational problem is shown to be equivalent to the abstract minimization problem. Let us confirm that in this section.

Let U be a real Hilbert space and $a : U \times U \rightarrow \mathbb{R}$ be a bilinear form on U . If for arbitrary $u, v \in U$,

$$a(u, v) = a(v, u)$$

holds, a is called symmetric.

If U is \mathbb{R}^d and with respect to $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $a(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot (\mathbf{A} \mathbf{y})$, a being symmetric is equivalent to the matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ being symmetric $\mathbf{A} = \mathbf{A}^\top$.

The following problem is called an abstract minimization problem.

Problem 5.2.8 (Abstract Minimization Problem) Let $a : U \times U \rightarrow \mathbb{R}$ be a bilinear form on U , $l = l(\cdot) = \langle l, \cdot \rangle \in U'$ and $f : U \rightarrow \mathbb{R}$. In this case, obtain $u \in U$ such that

$$\min_{u \in U} \left\{ f(u) = \frac{1}{2}a(u, u) - l(u) \right\}. \quad \square$$

If $U = \mathbb{R}^d$, it becomes a problem seeking $\mathbf{x} \in \mathbb{R}^d$ satisfying

$$\min_{\mathbf{x} \in \mathbb{R}^d} \left\{ f(\mathbf{x}) = \frac{1}{2}\mathbf{x} \cdot (\mathbf{A}\mathbf{x}) - \mathbf{b} \cdot \mathbf{x} \right\}. \quad (5.2.7)$$

The following results can be obtained with respect to Problem 5.2.8 (cf. [42, Theorem 1.1, p. 24], [95, Theorem 2.1, p. 33], [158, Theorem 2.7, p. 50]).

Theorem 5.2.9 (Solution to Abstract Minimization Problem) *Consider Problem 5.2.8 and let a be coercive, bounded and symmetric. In this case, with respect to an arbitrary $l \in U'$, $u \in U$ satisfying Problem 5.2.8 exists uniquely and agrees with the solution to Problem 5.2.3.* \square

If $U = \mathbb{R}^d$ and A is bounded, positive definite and symmetric, $x \in \mathbb{R}^d$ satisfying Eq. (5.2.7) is the same as Eq. (5.2.2).

If a is symmetric in the weak form of the Poisson problem (Problem 5.1.2), then Problem 5.1.2 is equivalent to the following problem in view of the solution of Exercise 5.2.5 and Theorem 5.2.9.

Problem 5.2.10 (Minimization Problem of Poisson Problem) Let a and \hat{l} be Eqs. (5.1.8) and (5.2.3), respectively. In this case, obtain $\tilde{u} = u - u_D \in U$ which satisfies

$$\min_{\tilde{u} \in U} \left\{ f(\tilde{u}) = \frac{1}{2}a(\tilde{u}, \tilde{u}) - \hat{l}(\tilde{u}) \right\}. \quad \square$$

5.3 Regularity of Solutions

The Poisson problem is an abstract variational problem and we have seen how the existence of a unique solution can be guaranteed by the Lax–Milgram theorem. In this case, if \hat{l} of Eq. (5.2.3) constructed from the given functions b , p , u_D in the Poisson problem (Problem 5.1.1) is in U' . This shows that the solution $u - u_D$ of the Poisson problem exists in $U = \{u \in H^1(\Omega; \mathbb{R}) \mid u = 0 \text{ on } \Gamma_D\}$. However, this condition is necessary for the existence of a solution. So, even if smoother given functions are assumed, it is expected that the solution to the Poisson problem will be correspondingly smooth. In Chaps. 8 and 9 smoothness greater than H^1 class is needed with respect to the solution to a boundary value problem. Here let us examine this notion of smoothness.

In this book, the smoothness of a function represents the order of differentiability and the exponent of integrability for the function. They are referred to as the regularity of function. In contrast, if there are not enough regularities or there are only a few, it is referred to as irregularity. The regularity (or irregularity) of a function can be expressed as “ C^1 class” by adding “class” to the symbol representing the function space.

There are two factors determining the irregularity of the solution to a boundary value problem. Let us look at these in the following subsections.

5.3.1 Regularity of Given Functions

Firstly, let us think about the relationship between the regularity of solution and regularity of given functions b , p , u_D of the Poisson problem (Problem 5.1.1). Suppose the boundary $\partial\Omega$ is sufficiently smooth. In this case, from the fact that

$$-\Delta u = b \quad \text{in } \Omega, \quad \partial_\nu u = p_N \quad \text{on } \Gamma_N, \quad u = u_D \quad \text{on } \Gamma_D,$$

holds, if we assume

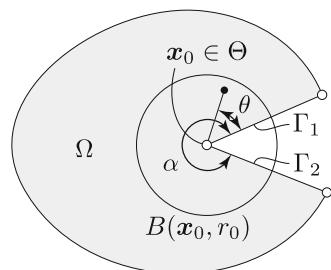
$$b \in L^2(\Omega; \mathbb{R}), \quad p_N \in H^1(\Omega; \mathbb{R}), \quad u_D \in H^2(\Omega; \mathbb{R}),$$

we get $u \in H^2(\Omega \setminus \bar{B}; \mathbb{R})$, where the neighborhood of the boundary between the Dirichlet boundary and the Neumann boundary is denoted by B . In more detail, if we set $b \in L^2(\Omega; \mathbb{R})$, we obtain $u \in H^2(\Omega \setminus \bar{B}; \mathbb{R})$ from the fact that the Poisson equation is satisfied. Moreover, since the boundary $\partial\Omega$ is sufficiently smooth, then $\nu \in C(\Gamma_N; \mathbb{R})$. So if $p_N \in H^1(\Omega; \mathbb{R})$, $\partial_\nu u = \nu \cdot \nabla u \in H^{1/2}(\Gamma_N \setminus \bar{B}; \mathbb{R})$ is obtained from $p_N \in H^{1/2}(\Gamma_N; \mathbb{R})$. From this, $u \in H^2(\Omega \setminus \bar{B}; \mathbb{R})$ can be obtained. Moreover, according to the Sobolev embedding theorem (Theorem 4.3.14), $H^2(\Omega; \mathbb{R}) \subset C^{0,\sigma}(\bar{\Omega}; \mathbb{R})$ holds with respect to $\sigma \in (0, 1/2)$ and $d \in \{2, 3\}$. Hence, u becomes a continuous function. It is said that there are no irregularities in the solution u in this case. If a given function is changed to an even smoother function then a correspondingly smoother u can be obtained.

5.3.2 Regularity of Boundary

On the other hand, even if given functions are assumed to be sufficiently smooth, if the boundary is not smooth, there can be irregularities in the solution. Let us look at such a situation in detail. In this section, Ω is assumed to be a two-dimensional domain and focus is given to the neighborhood around a corner such as x_0 in Fig. 5.2. Such a corner point corresponds to looking at a point in the perpendicular

Fig. 5.2 Two-dimensional domain with a corner



cross-section with respect to a smooth cut-out line in a three-dimensional domain with a V-shaped cut-out.

A discontinuous point on boundary $\partial\Omega$ with respect to C^1 class such as x_0 on Fig. 5.2 is called a corner point. A set of corner points will be denoted by Θ . Let r_0 be a positive constant and $B(x_0, r_0)$ the neighborhood (open set) around x_0 with radius r_0 . Set the opening angle at x_0 in the internal domain to be $\alpha \in (0, 2\pi)$. The boundaries (open set) on both sides of x_0 in $B(x_0, r_0)$ are set to be Γ_1 and Γ_2 , respectively. The boundaries Γ_1 and Γ_2 are assumed to be smooth (C^1 class). Moreover, the polar coordinates with x_0 as the origin are denoted as (r, θ) .

From the fact that u is smooth (analytic) at points further away from x_0 , if some $r \in (0, r_0]$ is fixed, u can be expanded as

$$u(r, \theta) = \sum_{i \in \{1, 2, \dots\}} k_i u_i(r) \tau_i(\theta) + u_R \quad (5.3.1)$$

(cf. [108], [155, Chap. 8, p. 257], [159], [56, Preface, p. ix, and Chap. 4, p. 182]). Here, u_R expresses the remainder term determined by the regularity of the given function. In contrast, the first term on the right-hand side of Eq. (5.3.1) arising due to the corner point is called the main term. For each $i \in \{1, 2, \dots\}$, the main terms k_i are real constants and $u_i(r)$ represent real-valued functions determined dependent on r . Moreover, $\tau_i(\theta)$ are determined in the following way by real-valued functions of $\theta \in (0, \alpha)$ dependent on boundary conditions. When Γ_1 and Γ_2 are both homogeneous Dirichlet boundaries ($u = 0$) and both homogeneous Neumann boundaries ($\partial_\nu u = 0$), these respectively become

$$\tau_i(\theta) = \sin \frac{i\pi}{\alpha} \theta, \quad (5.3.2)$$

$$\tau_i(\theta) = \cos \frac{i\pi}{\alpha} \theta. \quad (5.3.3)$$

In reality, Eq. (5.3.2) satisfies $\tau_i(0) = \tau_i(\alpha) = 0$. Equation (5.3.3) satisfies the condition that $(d\tau_i/d\theta)(0) = (d\tau_i/d\theta)(\alpha) = 0$. Moreover, if Γ_1 and Γ_2 are mixed boundaries with a homogeneous Dirichlet and Neumann boundary, it becomes

$$\tau_i(\theta) = \sin \frac{i\pi}{2\alpha} \theta. \quad (5.3.4)$$

On the other hand, with respect to the Laplace operator Δ ,

$$\Delta(r^\omega \sin \omega\theta) = \left(\frac{\partial^2}{\partial r^2} + \frac{1}{r} \frac{\partial}{\partial r} + \frac{1}{r^2} \frac{\partial^2}{\partial \theta^2} \right) (r^\omega \sin \omega\theta) = 0 \quad (5.3.5)$$

holds, where ω is a real number satisfying $\omega > 1/4$ which is not 1. The condition $\omega > 1/4$ corresponds to the fact that, in the condition shown later, as Γ_1 and Γ_2 have mixed boundary conditions and get closer to a crack ($\alpha \rightarrow 2\pi$), they become

$\omega \rightarrow 1/4$. Moreover, $\omega = 1$ corresponds to the condition that the boundary is smooth. In addition, Eq. (5.3.5) is also obtained by Cauchy–Riemann equations, which forms a necessary and sufficient condition for a complex function to be complex differentiable (holomorphic), with respect to the imaginary part $u_i = \operatorname{Im}[z^\omega] = r^m \sin \omega \theta$ of a complex function $f(z) = z^\omega$ using the correspondence between a complex number $z = x_1 + ix_2 = r e^{i\theta} \in \mathbb{C}$ (i is the imaginary unit) and $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$. Equation (5.3.5) shows that if a function has the format $r^\omega \sin \omega \theta$, the Laplace equation (same with the homogeneous Poisson equation) is satisfied.

Focusing on this relationship, when $\tau_i(\theta)$ is given in the format $\sin \omega \theta$, if

$$u_i(r) = r^\omega,$$

the Laplace equation is satisfied. From this result, the following results are obtained in the neighborhood $B(\mathbf{x}_0, r_0) \cap \Omega$ around \mathbf{x}_0 with radius r_0 :

(1) When Γ_1 and Γ_2 are both homogeneous Dirichlet boundaries ($u = 0$),

$$u(r, \theta) = kr^{\pi/\alpha} \sin \frac{\pi}{\alpha} \theta + u_R. \quad (5.3.6)$$

(2) When Γ_1 and Γ_2 are both homogeneous Neumann boundaries ($\partial_\nu u = 0$),

$$u(r, \theta) = kr^{\pi/\alpha} \cos \frac{\pi}{\alpha} \theta + u_R. \quad (5.3.7)$$

(3) If it is a mixed boundary where Γ_1 is a homogeneous Dirichlet boundary and Γ_2 is a homogeneous Neumann boundary,

$$u(r, \theta) = kr^{\pi/(2\alpha)} \sin \frac{\pi}{2\alpha} \theta + u_R. \quad (5.3.8)$$

Here, k is a constant dependent on α .

Moreover, the following result can be obtained for a Sobolev space containing functions of the format r^ω .

Proposition 5.3.1 (Regularity of Singularity Term) *Let Ω be a two-dimensional bounded domain and \mathbf{x}_0 be a corner point of opening angle $\alpha \in (0, 2\pi)$ on $\partial\Omega$. The function u is given by*

$$u = r^\omega \tau(\theta)$$

in the neighborhood $B(\mathbf{x}_0, r_0) \cap \Omega$ around \mathbf{x}_0 , where $\tau(\theta)$ is taken to be an element of $C^\infty((0, \alpha), \mathbb{R})$. In this case, if

$$\omega > k - \frac{2}{p} \quad (5.3.9)$$

holds for $k \in \mathbb{N} \cup \{0\}$ and $p \in (1, \infty)$, u is in $W^{k,p}(B(\mathbf{x}_0, r_0) \cap \Omega; \mathbb{R})$. □

Proof The k -th order derivative of $u = r^\omega \tau(\theta)$ is constructed as a sum of the terms including $r^{\omega-k} \tilde{\tau}(\theta)$. Here, $\tilde{\tau}(\theta)$ is an element of $C^\infty((0, \alpha), \mathbb{R})$. Hence, in order for the p -th-order Lebesgue integral on $B(\mathbf{x}_0, r_0) \cap \Omega$ of k -th order derivative of u to be finite, the condition

$$\int_0^{r_0} \int_0^\alpha r^{p(\omega-k)} r \tilde{\tau}(\theta) d\theta dr < \infty$$

needs to hold. For this,

$$p(\omega - k) + 1 > -1$$

is obtained. This relationship gives Eq. (5.3.9). \square

From the fact that the main term of solution u to the Poisson problem around the corner point is a function of the form r^ω and Proposition 5.3.1, the following results can be obtained with respect to a corner point such as that in Fig. 5.3.

Theorem 5.3.2 (Regularity of a Solution Around a Corner) *Let Ω be a two-dimensional bounded domain and $\mathbf{x}_0 \in \Theta$ be a corner point of opening angle $\alpha \in (0, 2\pi)$. In this case the solution u of the Poisson problem (Problem 5.1.1) is in $H^s(B(\mathbf{x}_0, r_0) \cap \Omega; \mathbb{R})$ in the neighborhood of \mathbf{x}_0 . Here:*

- (1) *if the boundaries Γ_1 and Γ_2 of both sides of \mathbf{x}_0 share the same type of boundary condition, then $\alpha \in [\pi, 2\pi)$ implies that $s \in (3/2, 2]$.*
- (2) *if Γ_1 and Γ_2 are mixed boundaries, then $\alpha \in [\pi/2, \pi)$ implies $s \in (3/2, 2]$ and $\alpha \in [\pi, 2\pi)$ means that $s \in (5/4, 3/2]$. \square*

Proof If Γ_1 and Γ_2 are the same type of boundary, Eqs. (5.3.6) and (5.3.7) give $\omega = \pi/\alpha$. Here, when the opening angle is $\alpha \in [\pi, 2\pi)$, $\omega \in (1/2, 1]$. In this case, if the inequality condition

$$s_1 - \frac{2}{p} = \frac{3}{2} - \frac{2}{2} = \frac{1}{2} < \omega \leq s_2 - \frac{2}{p} = 2 - \frac{2}{2} = 1$$

holds with respect to Eq. (5.3.9), then $\omega \in (1/2, 1]$ implies that s is (1).

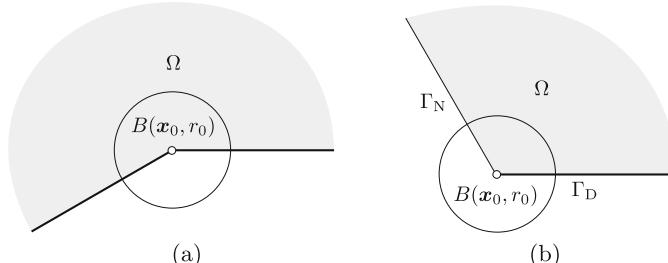


Fig. 5.3 Two-dimensional domain having a corner with irregularity. (a) Opening angle is $\alpha > \pi$ between boundaries of the same type. (b) Opening angle is $\alpha > \pi/2$ between the mixed boundaries

On the other hand, if Γ_1 and Γ_2 are mixed boundaries, Eq. (5.3.8) gives $\omega = \pi/(2\alpha)$. Hence, if the opening angle is $\alpha \in [\pi/2, \pi)$, then $\omega \in (1/2, 1]$ and s satisfying Eq. (5.3.9) becomes a result such as that in the first half of (2). Moreover, if the opening angle is $\alpha \in [\pi, 2\pi)$, then $\omega \in (1/4, 1/2]$. In this case, if we have that

$$s_1 - \frac{2}{p} = \frac{5}{4} - \frac{2}{2} = \frac{1}{4} < \omega \leq s_2 - \frac{2}{p} = \frac{3}{2} - \frac{2}{2} = \frac{1}{2}$$

holds with respect to Eq. (5.3.9), then s becomes a result such as the latter half of (2) with respect to $\omega \in (1/4, 1/2]$. \square

Assumptions in Theorem 5.3.2 did not include a crack ($\alpha = 2\pi$). If \mathbf{x}_0 is a crack tip,

$$u \in H^{3/2-\epsilon}(B(\mathbf{x}_0, r_0) \cap \Omega; \mathbb{R}) \quad (5.3.10)$$

can be written with respect to $\epsilon > 0$. Moreover, even when \mathbf{x}_0 is a boundary of mixed boundaries and the boundary is smooth around \mathbf{x}_0 ($\alpha = \pi$), it can be written as Eq. (5.3.10).

In order to guarantee that u is a function of $W^{1,\infty}$ class, the following result can be used.

Theorem 5.3.3 (Regularity of a Solution Around a Corner) *Let Ω be a two-dimensional bounded domain and $\mathbf{x}_0 \in \Theta$ be a corner point of opening angle $\alpha \in (0, 2\pi)$. The solution u of the Poisson problem (Problem 5.1.1) is in $W^{1,\infty}(B(\mathbf{x}_0, r_0) \cap \Omega; \mathbb{R})$,*

- (1) *if $\alpha < \pi$ in the case that the boundaries Γ_1 and Γ_2 of both sides of \mathbf{x}_0 share the same type of boundary condition,*
- (2) *if $\alpha < \pi/2$ in the case that Γ_1 and Γ_2 are mixed boundaries.* \square

Proof If Γ_1 and Γ_2 are the same type of boundary, Eqs. (5.3.6) and (5.3.7) give $\omega = \pi/\alpha$. Here, when the opening angle is $\alpha < \pi$, $\omega > 1$. In this case, (1) holds with respect to Eq. (5.3.9). On the other hand, if Γ_1 and Γ_2 are mixed boundaries, Eq. (5.3.8) gives $\omega = \pi/(2\alpha)$. Hence, if the opening angle is $\alpha < \pi/2$, then $\omega > 1$. From Eq. (5.3.9), we have that (2) holds with respect to Eq. (5.3.9). \square

Moreover, from Theorem 5.3.2 (2), if Γ_1 and Γ_2 are mixed boundaries, it becomes apparent that even when the boundary is smooth, the same irregularity is observed as that at a crack tip. One method for preventing the occurrence of such irregularity is to rewrite the mixed boundary value problem as an extended Poisson problem, such as Problem 5.1.3. In this case, by assuming a smooth function in $c_{\partial\Omega} : \partial\Omega \rightarrow \mathbb{R}$ such that it changes from a Dirichlet boundary to a Neumann boundary, a mixed boundary value problem with no singularities can be constructed.

5.4 Linear Elastic Problem

In this book, attempts are made to represent specific examples of shape optimization problems using a linear elastic body and Stokes flow field. As preparation for that, we shall now define a linear elastic problem and look at the existence of unique solutions and their weak forms.

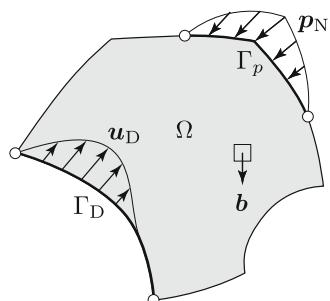
Let $\Omega \subset \mathbb{R}^d$ be a $d \in \{2, 3\}$ -dimensional Lipschitz domain. Let $\Gamma_D \subset \partial\Omega$ be a boundary when displacement is given (Dirichlet boundary) and the remaining boundary $\Gamma_N = \partial\Omega \setminus \bar{\Gamma}_D$ be a boundary where traction is given (Neumann boundary). Moreover, $\Gamma_p \subset \Gamma_N$ is taken to represent a boundary where the traction is non-zero. Here, Γ_p and $\Gamma_N \setminus \bar{\Gamma}_p$ are not distinguished but they will be in Chap. 9. Figure 5.4 shows a linear elastic body in the two-dimensional case. However, as seen in Exercise 5.2.6, in order to get rid of the uncertainty of constant, $|\Gamma_D| > 0$ is assumed. Moreover, $\mathbf{b} : \Omega \rightarrow \mathbb{R}^d$ is taken to be a volume force, $\mathbf{p}_N : \Gamma_N \rightarrow \mathbb{R}^d$ is the traction and $\mathbf{u}_D : \Omega \rightarrow \mathbb{R}^d$ is the given displacement. A linear elastic problem is defined as a problem seeking displacements $\mathbf{u} : \Omega \rightarrow \mathbb{R}^d$ when these are given.

5.4.1 Linear Strain

A linear elastic problem of a one-dimensional continuous body was defined in Chap. 1. Here, let us extend this to $d \in \{2, 3\}$ dimensions. Firstly, let us define the term strain. In a one-dimensional linear elastic body, the displacement u was a real-valued function defined on $(0, l)$. Strain was defined using its gradient du/dx . If the linear elastic body is $d \in \{2, 3\}$ -dimensional, the displacement \mathbf{u} becomes a d -dimensional vector and its gradient $(\nabla \mathbf{u}^\top)^\top = (\partial u_i / \partial x_j)_{ij}$ becomes a second-order tensor (matrix) with the value $\mathbb{R}^{d \times d}$. Figure 5.5 shows the relationship between \mathbf{u} and $(\nabla \mathbf{u}^\top)^\top$. This tensor is split into the symmetric and non-symmetric components as

$$(\nabla \mathbf{u}^\top)^\top = \mathbf{E}(\mathbf{u}) + \mathbf{R}(\mathbf{u}) \quad (5.4.1)$$

Fig. 5.4 Two-dimensional linear elastic problem



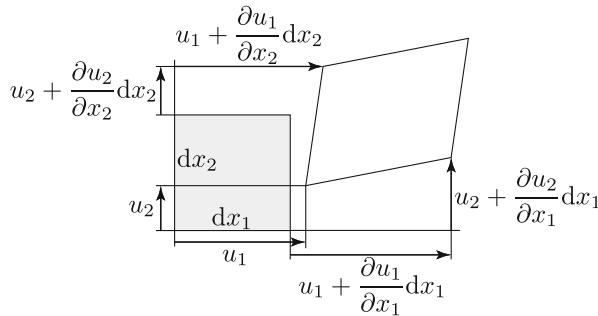


Fig. 5.5 Displacement \mathbf{u} and its gradient $(\nabla \mathbf{u}^\top)^\top$ in 2D linear elastic body

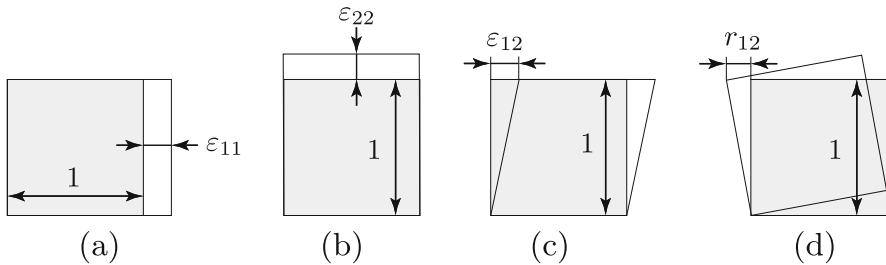


Fig. 5.6 Linear strain $\mathbf{E}(\mathbf{u})$ and rotation tensor $\mathbf{R}(\mathbf{u})$ in 2D linear elastic body. (a) ε_{11} . (b) ε_{22} . (c) $\varepsilon_{12} = \varepsilon_{21}$. (d) r_{12}

In this case,

$$\mathbf{E}(\mathbf{u}) = \mathbf{E}^\top(\mathbf{u}) = (\varepsilon_{ij}(\mathbf{u}))_{ij} = \frac{1}{2} \left(\nabla \mathbf{u}^\top + (\nabla \mathbf{u}^\top)^\top \right), \quad (5.4.2)$$

$$\mathbf{R}(\mathbf{u}) = -\mathbf{R}^\top(\mathbf{u}) = (r_{ij}(\mathbf{u}))_{ij} = \frac{1}{2} \left((\nabla \mathbf{u}^\top)^\top - \nabla \mathbf{u}^\top \right). \quad (5.4.3)$$

Here, the symmetric component $\mathbf{E}(\mathbf{u})$ represents the deformations such as those from (a) to (c) in Fig. 5.6 when Ω is a two-dimensional domain, and is called the linear strain, or simply strain of a d -dimensional linear elastic body if there is no confusion. Moreover, the non-symmetric component $\mathbf{R}(\mathbf{u})$ represents the rotational motion such as (d) in Fig. 5.6 with respect to a two-dimensional domain Ω , and is called the rotation tensor of a d -dimensional linear elastic body.

The linear strain and rotation tensor defined in Eqs. (5.4.2) and (5.4.3) were defined using the gradient tensor of \mathbf{u} when \mathbf{u} is $\mathbf{0}_{\mathbb{R}^d}$ (before deformation). Hence, there is a need to focus on the fact that \mathbf{u} cannot take a large value. When it is assumed that \mathbf{u} is finite, finite deformation theory using Green strain or Almansi strain which is defined with the second-order terms of elements of the gradient tensor of displacement is employed. In this case the differential equation becomes

non-linear. The non-linearity in this case is called a geometric non-linearity. This book is limited to linear problems.

5.4.2 Cauchy Tensor

In contrast, with respect to a linear strain defined from displacement, stress can be defined from the distribution of force. Consider a small domain inside the domain Ω . When $d = 2$, a triangle such as the one in Fig. 5.7b is imagined, while when $d = 3$, a triangular pyramid such as that in Fig. 5.8 is considered. The normal of their tilt boundary is ν . Force per unit boundary measure (length when $d = 2$, area when $d = 3$) working on the tilt boundary is taken to be $\mathbf{p} \in \mathbb{R}^d$. The function \mathbf{p} represents the stress. Moreover, with respect to $i, j \in \{1, \dots, d\}$, when σ_{ij} is the force in the x_j -direction per unit boundary measure working on a boundary with normal in the x_i -direction, $\mathbf{S} = (\sigma_{ij}) \in \mathbb{R}^{d \times d}$ is called Cauchy stress, or simply stress if there is no confusion.

Cauchy stress \mathbf{S} and stress \mathbf{p} can be related in the following way.

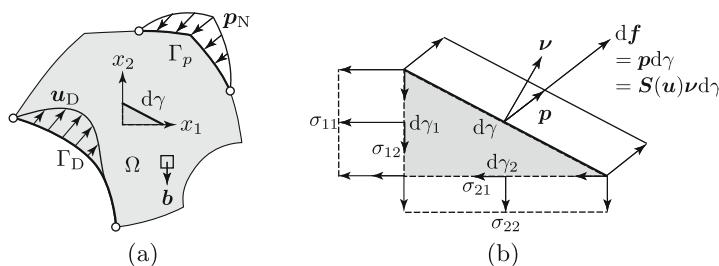


Fig. 5.7 Cauchy stress \mathbf{S} and stress \mathbf{p} of 2D linear elastic body. (a) Small line component $d\gamma$ in domain, (b) Cauchy stress and stress

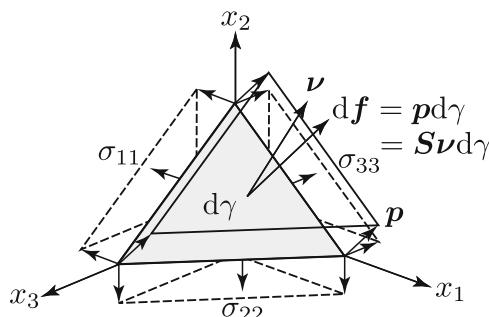


Fig. 5.8 Cauchy stress \mathbf{S} and stress \mathbf{p} of 3D elastic body

Proposition 5.4.1 (Cauchy Tensor) When \mathbf{p} is a stress and \mathbf{S} is its Cauchy stress,

$$\mathbf{S}^\top \mathbf{v} = \mathbf{S}\mathbf{v} = \mathbf{p} \quad (5.4.4)$$

holds. \square

Proof We will show the case when $d = 2$. From the balance of force in the direction x_i with respect to $i \in \{1, 2\}$, the relation

$$\sigma_{1i} d\gamma_1 + \sigma_{2i} d\gamma_2 = p_i d\gamma$$

holds (Fig. 5.7b). Here, using $\nu_1 = d\gamma_1/d\gamma$ and $\nu_2 = d\gamma_2/d\gamma$ gives

$$\sigma_{1i} \nu_1 + \sigma_{2i} \nu_2 = p_i. \quad (5.4.5)$$

Equation (5.4.5) represents Eq. (5.4.4). On the other hand, the identity

$$\sigma_{21} = \sigma_{12}$$

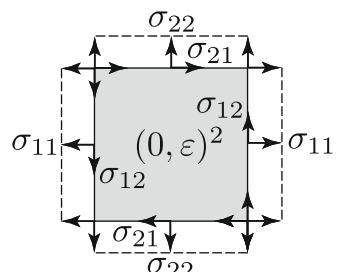
holds from the balance of moments (Fig. 5.9). A similar relationship holds when $d = 3$ too. \square

5.4.3 Constitutive Equation

In a similar way to that used in Chap. 1 when one-dimensional linear elastic problems were defined, constitutive equation or constitutive law, which relates the strain defined using displacement and the stress defined using force, is now needed. In a linear elastic body of d dimensions, it is given by

$$\begin{aligned} \mathbf{S}(\mathbf{u}) &= \mathbf{S}^\top(\mathbf{u}) = (\sigma_{ij}(\mathbf{u}))_{ij} \\ &= \mathbf{C}\mathbf{E}(\mathbf{u}) = \left(\sum_{(k,l) \in \{1, \dots, d\}^2} c_{ijkl} \varepsilon_{kl}(\mathbf{u}) \right)_{ij}. \end{aligned} \quad (5.4.6)$$

Fig. 5.9 Balance of moments at a small area in a two-dimensional linear elastic body ($\epsilon \ll 1$)



Here, $\mathbf{C} = (c_{ijkl})_{ijkl} : \Omega \rightarrow \mathbb{R}^{d \times d \times d \times d}$ is a function of fourth-order tensor value representing the rigidity and assumes the following characteristics. Firstly, from the symmetry of $\mathbf{S}(\mathbf{u})$ and $\mathbf{E}(\mathbf{u})$, the relationships

$$c_{ijkl} = c_{jikl}, \quad c_{ijkl} = c_{ijlk}. \quad (5.4.7)$$

hold. Moreover, assuming \mathbf{C} is L^∞ class, there exist positive constants α and β such that

$$\mathbf{A} \cdot (\mathbf{C}\mathbf{A}) \geq \alpha \|\mathbf{A}\|^2, \quad (5.4.8)$$

$$|\mathbf{A} \cdot (\mathbf{C}\mathbf{B})| \leq \beta \|\mathbf{A}\| \|\mathbf{B}\| \quad (5.4.9)$$

hold almost everywhere in Ω with respect to arbitrary symmetric tensor $\mathbf{A} = (a_{ij})_{ij} \in \mathbb{R}^{d \times d}$ and $\mathbf{B} = (b_{ij})_{ij} \in \mathbb{R}^{d \times d}$. In this book, the scalar product of matrices is represented as $\mathbf{A} \cdot \mathbf{B} = \sum_{i,j \in \{1, \dots, d\}} a_{ij} b_{ij}$. The fact that Eq. (5.4.8) holds is referred to as \mathbf{C} being elliptic. Moreover, the fact that Eq. (5.4.9) holds is referred to as \mathbf{C} being bounded. When \mathbf{C} is not a function of \mathbf{u} (stress is a linear function of strain), Eq. (5.4.6) is referred to as the generalized Hooke's law. The non-linearity such that \mathbf{C} becomes a function of \mathbf{u} is called material non-linearity. This sort of non-linearity will also not be treated in this book.

Moreover, the following can be said about the number of real numbers which can be chosen independently in the rigidity \mathbf{C} , when $d = 3$:

- (1) \mathbf{C} is constructed of $3^4 = 81$ real numbers.
- (2) It reduces to 36 from Eq. (5.4.7).
- (3) If strain energy density w exists and

$$w = \frac{1}{2} \mathbf{E}(\mathbf{u}) \cdot (\mathbf{C}\mathbf{E}(\mathbf{u})), \quad \mathbf{S}(\mathbf{u}) = \frac{\partial w}{\partial \mathbf{E}(\mathbf{u})}$$

holds, the relation

$$c_{ijkl} = c_{klji} \quad (5.4.10)$$

holds by the symmetry of second-order form. In this case it reduces down to 21.

- (4) It reduces to nine in the case of orthotropic materials.
- (5) It reduces to two in the case of isotropic materials.

Suppose the two constants in the case of isotropic material are written as λ_L and μ_L and

$$\mathbf{S}(\mathbf{u}) = 2\mu_L \mathbf{E}(\mathbf{u}) + \lambda_L \text{tr}(\mathbf{E}(\mathbf{u})) \mathbf{I},$$

where $\text{tr}(\mathbf{E}(\mathbf{u})) = \sum_{i \in \{1, \dots, d\}} e_{ii}(\mathbf{u})$. In this case, the quantities λ_L and μ_L are called Lamé's parameters. Moreover, μ_L is also referred to as shear modulus. In

addition, when the two constants are expressed as e_Y and ν_P and

$$\mathbf{E}(\mathbf{u}) = \frac{1 + \nu_P}{e_Y} \mathbf{S}(\mathbf{u}) - \frac{\nu_P}{e_Y} \text{tr}(\mathbf{S}(\mathbf{u})) \mathbf{I}$$

is assumed, e_Y and ν_P are called longitudinal elastic modulus (Young's modulus) and Poisson's ratio, respectively. Other than this, bulk modulus k_b is also used. A relationship such as

$$k_b = \lambda_L + \frac{2\mu_L}{3}, \quad e_Y = 2\mu_L(1 + \nu_P), \quad \lambda_L = \frac{2\mu_L \nu_P}{1 - 2\nu_P}$$

holds with respect to these constants.

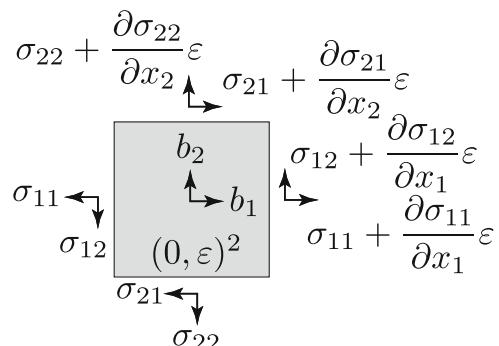
5.4.4 Equilibrium Equations of Force

A linear elastic problem is constructed using the balance condition of forces based on a linear strain and Cauchy stress being linked via generalized Hooke's law Eq. (5.4.6).

When an arbitrary small square element is chosen within a two-dimensional linear elastic body, the force working on that element is as shown by the arrows in Fig. 5.10. Here, the equilibrium equation of force in the x_1 -direction and x_2 -direction becomes

$$\begin{aligned} \frac{\partial \sigma_{11}}{\partial x_1} + \frac{\partial \sigma_{21}}{\partial x_2} + b_1 &= 0, \\ \frac{\partial \sigma_{12}}{\partial x_1} + \frac{\partial \sigma_{22}}{\partial x_2} + b_2 &= 0. \end{aligned}$$

Fig. 5.10 Balance of forces in a small area ($\epsilon \ll 1$)



In the case of a $d \in \{2, 3\}$ -dimensional linear elastic body, it can be written as

$$-\nabla^T S(\mathbf{u}) = \mathbf{b}^T. \quad (5.4.11)$$

Equation (5.4.11) is a second-order differential equation with respect to \mathbf{u} if we look at the fact that $\nabla^T S(\mathbf{u}) = \nabla \cdot \left\{ \mathbf{C} \left(\frac{1}{2} (\nabla \mathbf{u}^T + (\nabla \mathbf{u}^T)^T) \right) \right\}$. Furthermore, from the fact that \mathbf{C} satisfies ellipticity, Eq. (5.4.11) is classed as an elliptic partial differential equation.

Adding boundary conditions to the equilibrium equation (Eq. (5.4.11)) of force gives a linear elastic problem such as the one below.

Problem 5.4.2 (Linear Elastic Problem) Let the functions $\mathbf{b} : \Omega \rightarrow \mathbb{R}^d$, $\mathbf{p}_N : \Gamma_N \rightarrow \mathbb{R}^d$ and $\mathbf{u}_D : \Omega \rightarrow \mathbb{R}^d$ be given. Obtain $\mathbf{u} : \Omega \rightarrow \mathbb{R}^d$ which satisfies

$$-\nabla^T S(\mathbf{u}) = \mathbf{b}^T \quad \text{in } \Omega, \quad (5.4.12)$$

$$S(\mathbf{u}) \mathbf{v} = \mathbf{p}_N \quad \text{on } \Gamma_N, \quad (5.4.13)$$

$$\mathbf{u} = \mathbf{u}_D \quad \text{on } \Gamma_D. \quad (5.4.14)$$

□

5.4.5 Weak Form

In order to show the existence of a unique solution to the linear elastic problem, let us rewrite Problem 5.4.2 in the weak form. Let the function space with respect to \mathbf{u} be

$$U = \left\{ \mathbf{v} \in H^1(\Omega; \mathbb{R}^d) \mid \mathbf{v} = \mathbf{0}_{\mathbb{R}^d} \text{ on } \Gamma_D \right\}. \quad (5.4.15)$$

By multiplying both sides of Eq. (5.4.12) by an arbitrary $\mathbf{v} \in U$ and integrating over Ω , then using the Gauss–Green theorem (Theorem A.8.2), the equation

$$\begin{aligned} - \int_{\Omega} (\nabla^T S(\mathbf{u})) \mathbf{v} \, dx &= - \int_{\Gamma_N} (S(\mathbf{u}) \mathbf{v}) \cdot \mathbf{v} \, d\gamma + \int_{\Omega} S(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}) \, dx \\ &= \int_{\Omega} \mathbf{b} \cdot \mathbf{v} \, dx \end{aligned} \quad (5.4.16)$$

can be obtained. Moreover, if both sides of Eq. (5.4.13) are multiplied by an arbitrary $\mathbf{v} \in U$ and integrated over Γ_N , the equation

$$\int_{\Gamma_N} (S(\mathbf{u}) \mathbf{v}) \cdot \mathbf{v} \, d\gamma = \int_{\Gamma_N} \mathbf{p}_N \cdot \mathbf{v} \, d\gamma \quad (5.4.17)$$

is obtained. Substituting Eq. (5.4.17) in the first term of the second equation in Eq. (5.4.16) gives

$$\int_{\Omega} S(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}) \, dx = \int_{\Omega} \mathbf{b} \cdot \mathbf{v} \, dx + \int_{\Gamma_N} \mathbf{p}_N \cdot \mathbf{v} \, d\gamma.$$

This equation, which holds for arbitrary $\mathbf{v} \in U$, is referred to as the weak form of the linear elastic problem.

Also, if we set

$$a(\mathbf{u}, \mathbf{v}) = \int_{\Omega} S(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}) \, dx, \quad (5.4.18)$$

$$l(\mathbf{v}) = \int_{\Omega} \mathbf{b} \cdot \mathbf{v} \, dx + \int_{\Gamma_N} \mathbf{p}_N \cdot \mathbf{v} \, d\gamma, \quad (5.4.19)$$

the weak-form linear elastic problem becomes as follows.

Problem 5.4.3 (Weak Form of Linear Elastic Problem) Let U be given by Eq. (5.4.15) and the functions $\mathbf{b} \in L^2(\Omega; \mathbb{R}^d)$, $\mathbf{p}_N \in L^2(\Gamma_N; \mathbb{R}^d)$, $\mathbf{u}_D \in H^1(\Omega; \mathbb{R}^d)$ and $\mathbf{C} \in L^\infty(\Omega; \mathbb{R}^{d \times d \times d \times d})$. Set $a(\cdot, \cdot)$ and $l(\cdot)$ as Eqs. (5.4.18) and (5.4.19), respectively. In this case, seek $\tilde{\mathbf{u}} = \mathbf{u} - \mathbf{u}_D \in U$ which satisfies

$$a(\mathbf{u}, \mathbf{v}) = l(\mathbf{v})$$

with respect to an arbitrary $\mathbf{v} \in U$. □

5.4.6 Existence of Solution

If \mathbf{v} is viewed as a virtual displacement, from the fact that $l(\mathbf{v})$ is the virtual work done by external forces and $a(\mathbf{u}, \mathbf{v})$ is the virtual work done by internal forces, the weak form of a linear elastic problem represents the principle of virtual work. The existence of unique solutions with respect to this weak form is shown as follows.

Exercise 5.4.4 (Existence of Unique Solution to Linear Elastic Problem) In Problem 5.4.3, show that the solution $\tilde{\mathbf{u}} = \mathbf{u} - \mathbf{u}_D \in U$ exists uniquely for $|\Gamma_D| > 0$. □

Answer Let us confirm that the assumptions of the Lax–Milgram theorem hold. Let U be a Hilbert space and let

$$\hat{l}(\mathbf{v}) = l(\mathbf{v}) - a(\mathbf{u}_D, \mathbf{v})$$

with respect to an arbitrary $\mathbf{v} \in U$. Problem 5.4.3 can be rewritten as the problem seeking $\tilde{\mathbf{u}} = \mathbf{u} - \mathbf{u}_D \in U$ which satisfies

$$a(\tilde{\mathbf{u}}, \mathbf{v}) = \hat{l}(\mathbf{v}).$$

Under these assumptions, the fact that the Lax–Milgram theorem holds can be confirmed in the following ways:

- (1) a is coercive. In fact, the rigid motion is not generated because of $|\Gamma_D| > 0$. Therefore, from Korn's second inequality (Theorem A.9.6), the estimate

$$\|\mathbf{v}\|_{H^1(\Omega; \mathbb{R}^d)}^2 \leq c \|\mathbf{E}(\mathbf{v})\|_{L^2(\Omega; \mathbb{R}^{d \times d})}^2$$

holds with respect to a positive constant c . From ellipticity of \mathbf{C} due to Eq. (5.4.8), the inequality

$$\begin{aligned} a(\mathbf{v}, \mathbf{v}) &= \int_{\Omega} \mathbf{E}(\mathbf{v}) \cdot (\mathbf{C} \mathbf{E}(\mathbf{v})) \, dx \\ &\geq c_1 \|\mathbf{E}(\mathbf{v})\|_{L^2(\Omega; \mathbb{R}^{d \times d})}^2 \geq \frac{c_1}{c} \|\mathbf{v}\|_{H^1(\Omega; \mathbb{R}^d)}^2 \end{aligned}$$

holds with respect to $\mathbf{v} \in U$. Here, c_1 is a positive constant multiplying together α of Eq. (5.4.8) and $|\Omega|$. If c_1/c is reset to be α , from Definition 5.2.1, a is coercive.

- (2) a is bounded. In fact, if the positive constant multiplying β of Eq. (5.4.9) and $|\Omega|$ is replaced by β , Definition 5.2.2 confirms the boundedness of a .
(3) $\hat{l} \in U'$. In fact, since $\partial\Omega$ assumes a Lipschitz boundary, the norm $\|\gamma\|_{\mathcal{L}(H^1(\Omega; \mathbb{R}^d); H^{1/2}(\partial\Omega; \mathbb{R}^d))}$ of the trace operator (Theorem 4.4.2) is bounded. Let this be $c_2 > 0$. Moreover, using Hölder's inequality, the following result holds:

$$\begin{aligned} |\hat{l}(\mathbf{v})| &\leq \int_{\Omega} |\mathbf{b} \cdot \mathbf{v}| \, dx + \int_{\Gamma_N} |\mathbf{p}_N \cdot \mathbf{v}| \, dy + \int_{\Omega} \beta |\mathbf{E}(\mathbf{u}_D) \cdot \mathbf{E}(\mathbf{v})| \, dx \\ &\leq \|\mathbf{b}\|_{L^2(\Omega; \mathbb{R}^d)} \|\mathbf{v}\|_{L^2(\Omega; \mathbb{R}^d)} + \|\mathbf{p}_N\|_{L^2(\Gamma_N; \mathbb{R}^d)} \|\mathbf{v}\|_{L^2(\Gamma_N; \mathbb{R}^d)} \\ &\quad + \beta \|\mathbf{E}(\mathbf{u}_D)\|_{L^2(\Omega; \mathbb{R}^{d \times d})} \|\mathbf{E}(\mathbf{v})\|_{L^2(\Omega; \mathbb{R}^{d \times d})} \\ &\leq \left(\|\mathbf{b}\|_{L^2(\Omega; \mathbb{R}^d)} + c_2 \|\mathbf{p}_N\|_{L^2(\Gamma_N; \mathbb{R}^d)} \right. \\ &\quad \left. + \beta \|\mathbf{E}(\mathbf{u}_D)\|_{L^2(\Omega; \mathbb{R}^{d \times d})} \right) \|\mathbf{v}\|_{H^1(\Omega; \mathbb{R}^d)}. \end{aligned}$$

Therefore, from the Lax–Milgram theorem there is a unique $\tilde{\mathbf{u}} = \mathbf{u} - \mathbf{u}_D \in U$ which satisfies Problem 5.4.3. \square

5.5 Stokes Problem

Next, let us define a Stokes problem as an example of a flow field and look at its weak form and the existence of a unique solution. A Stokes problem is used as a mathematical model of a flow field which is slow so that inertia can be ignored relative to viscosity in the flow field of a viscous fluid.

In this case too, $\Omega \subset \mathbb{R}^d$ is taken to be a Lipschitz domain of $d \in \{2, 3\}$ dimensions. Again, let $\mathbf{b} : \Omega \rightarrow \mathbb{R}^d$ be the volume force. Let the entire boundary $\partial\Omega$ of Ω be a Dirichlet boundary with respect to flow velocity given by $\mathbf{u}_D : \Omega \rightarrow \mathbb{R}^d$ such that

$$\nabla \cdot \mathbf{u}_D = 0 \quad \text{in } \Omega. \quad (5.5.1)$$

Let μ be a positive constant representing coefficient of viscosity. Figure 5.11 shows a Stokes problem in two dimensions.

When these assumptions are given, a Stokes problem can be defined as a problem seeking flow velocity $\mathbf{u} : \Omega \rightarrow \mathbb{R}^d$ and pressure $p : \Omega \rightarrow \mathbb{R}$ in the following way. Here, $(\mathbf{v} \cdot \nabla) \mathbf{u} = (\nabla \mathbf{u}^\top)^\top \mathbf{v}$ is written as $\partial_\mathbf{v} \mathbf{u}$.

Problem 5.5.1 (Stokes Problem) Let $\mathbf{b} : \Omega \rightarrow \mathbb{R}^d$, $\mathbf{u}_D : \Omega \rightarrow \mathbb{R}^d$ and $\mu \in \mathbb{R}$ be given. Find $(\mathbf{u}, p) : \Omega \rightarrow \mathbb{R}^{d+1}$ such that the following equations,

$$-\nabla^\top (\mu \nabla \mathbf{u}^\top) + \nabla^\top p = \mathbf{b}^\top \quad \text{in } \Omega, \quad (5.5.2)$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega, \quad (5.5.3)$$

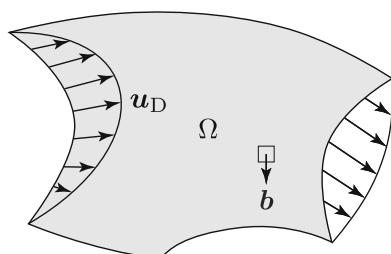
$$\mathbf{u} = \mathbf{u}_D \quad \text{on } \partial\Omega, \quad (5.5.4)$$

$$\int_{\Omega} p \, dx = 0, \quad (5.5.5)$$

are satisfied. \square

In Problem 5.5.1, Eq. (5.5.2) is called a Stokes equation and Eq. (5.5.3) is called a continuity equation. These are used to model the flow field of an incompressible fluid with the Newton viscosity.

Fig. 5.11 Two-dimensional Stokes problem



Moreover, Eq. (5.5.2) can be written as

$$-\nabla^\top (\mu \nabla \mathbf{u}^\top - p \mathbf{I}) = \mathbf{b}^\top \quad \text{in } \Omega, \quad (5.5.6)$$

where \mathbf{I} represents a unit matrix of d -th order. Moreover, defining Cauchy stress as

$$\mathbf{S}(\mathbf{u}, p) = -p \mathbf{I} + 2\mu \mathbf{E}(\mathbf{u}) \quad (5.5.7)$$

using $\mathbf{E}(\mathbf{u})$ defined in Eq. (5.4.2), Eq. (5.5.2) can be written as

$$-\nabla^\top \mathbf{S}(\mathbf{u}, p) = \mathbf{b}^\top \quad \text{in } \Omega. \quad (5.5.8)$$

If Eq. (5.5.3) holds, these are equivalent to one another. In this chapter, Eq. (5.5.2) is used in order to look at the relationship with the abstract saddle point variational problem in Sect. 5.6.

The weak form with respect to Problem 5.5.1 can be obtained in the following way. Let the function space with respect to \mathbf{u} be

$$U = H_0^1(\Omega; \mathbb{R}^d) = \left\{ \mathbf{u} \in H^1(\Omega; \mathbb{R}^d) \mid \mathbf{u} = \mathbf{0}_{\mathbb{R}^d} \text{ on } \partial\Omega \right\}. \quad (5.5.9)$$

Multiplying both sides of Eq. (5.5.2) by an arbitrary $\mathbf{v} \in U$ and integrating over Ω , then using the Gauss–Green theorem (Theorem A.8.2) gives

$$\begin{aligned} & \int_{\Omega} \left\{ \nabla^\top (\mu \nabla \mathbf{u}^\top) - \nabla^\top p + \mathbf{b}^\top \right\} \mathbf{v} \, dx \\ &= \int_{\partial\Omega} (\mu \partial_{\mathbf{v}} \mathbf{u} - p \mathbf{v}) \cdot \mathbf{v} \, d\gamma \\ & \quad + \int_{\Omega} \left(-\mu (\nabla \mathbf{u}^\top) \cdot (\nabla \mathbf{v}^\top) + p \nabla \cdot \mathbf{v} + \mathbf{b} \cdot \mathbf{v} \right) \, dx \\ &= \int_{\Omega} \left(-\mu (\nabla \mathbf{u}^\top) \cdot (\nabla \mathbf{v}^\top) + p \nabla \cdot \mathbf{v} + \mathbf{b} \cdot \mathbf{v} \right) \, dx \\ &= 0. \end{aligned}$$

The fact that this equation holds with respect to an arbitrary $\mathbf{v} \in U$ is referred to as the weak form of the Stokes equation.

On the other hand, let the function space with respect to p be

$$P = \left\{ q \in L^2(\Omega; \mathbb{R}) \mid \int_{\Omega} q \, dx = 0 \right\}. \quad (5.5.10)$$

Multiplying Eq. (5.5.3) by an arbitrary $q \in P$ and integrating over Ω gives

$$\int_{\Omega} q \nabla \cdot \mathbf{u} \, dx = 0.$$

The fact that this equation holds with respect to an arbitrary $q \in P$ is called the weak form of the continuity equation.

With respect to the Stokes problem, let

$$a(\mathbf{u}, \mathbf{v}) = \int_{\Omega} \mu (\nabla \mathbf{u}^{\top}) \cdot (\nabla \mathbf{v}^{\top}) \, dx, \quad (5.5.11)$$

$$b(\mathbf{v}, q) = - \int_{\Omega} q \nabla \cdot \mathbf{v} \, dx, \quad (5.5.12)$$

$$l(\mathbf{v}) = \int_{\Omega} \mathbf{b} \cdot \mathbf{v} \, dx. \quad (5.5.13)$$

In this case, the weak form of the Stokes problem can be written as follows.

Problem 5.5.2 (Weak Form of Stokes Problem) Let U and P be given by Eqs. (5.5.9) and (5.5.10), respectively. Suppose $\mathbf{u}_D \in H^1(\Omega; \mathbb{R}^d)$ satisfies Eq. (5.5.1). Let μ be a positive constant and $a(\cdot, \cdot)$, $b(\cdot, \cdot)$ and $l(\cdot)$ are taken to be Eqs. (5.5.11), (5.5.12) and (5.5.13), respectively. In this case, find $(\tilde{\mathbf{u}}, p) = (\mathbf{u} - \mathbf{u}_D, p) \in U \times P$ such that

$$a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) = l(\mathbf{v}), \quad (5.5.14)$$

$$b(\mathbf{u}, q) = 0, \quad (5.5.15)$$

for an arbitrary $(\mathbf{v}, q) \in U \times P$. □

5.6 Abstract Saddle Point Variational Problem

Now we have the weak form of the Stokes problem, let us look at what assumptions guarantee the existence of a unique solution of the given weak form.

A linear elastic problem is an elliptic partial differential equation with respect to displacement \mathbf{u} . Hence, the existence of a unique solution could be shown using the results with respect to an abstract variational problem or an abstract minimization problem. In contrast, a Stokes problem demands that the pressure p is added as an unknown variable in addition to the flow velocity \mathbf{u} , and that the continuity equation is satisfied simultaneously. This structure can be seen to be a problem called the abstract saddle point variational problem or abstract saddle point problem corresponding to the abstract variational problem with constraints or abstract minimization problem with constraints. Here, let us show the existence of

a unique solution with respect to the Stokes problem using certain definitions and results.

Let U and P be real Hilbert spaces, and the functions $a : U \times U \rightarrow \mathbb{R}$ and $b : U \times P \rightarrow \mathbb{R}$ be bounded bilinear operators defined on $U \times U$ and $U \times P$, respectively (Sect. 4.4.4). Also, let their norms be given by

$$\|a\| = \|a\|_{\mathcal{L}(U, U; \mathbb{R})} = \sup_{\mathbf{u}, \mathbf{v} \in U \setminus \{\mathbf{0}_U\}} \frac{|a(\mathbf{u}, \mathbf{v})|}{\|\mathbf{u}\|_U \|\mathbf{v}\|_U},$$

$$\|b\| = \|b\|_{\mathcal{L}(U, P; \mathbb{R})} = \sup_{\mathbf{u} \in U \setminus \{\mathbf{0}_U\}, q \in P \setminus \{\mathbf{0}_P\}} \frac{|b(\mathbf{u}, q)|}{\|\mathbf{u}\|_U \|q\|_P},$$

respectively.

A set of functions which satisfy the continuity equation that become Hilbert spaces is defined as follows.

Definition 5.6.1 (Divergence Free Hilbert Space U_{div}) Let $b : U \times P \rightarrow \mathbb{R}$ be a bilinear form. In this case,

$$U_{\text{div}} = \{\mathbf{v} \in U \mid b(\mathbf{v}, q) = 0 \text{ for all } q \in P\}$$

is called a divergence free Hilbert space of U . □

Using these definitions, we consider the following problem.

Problem 5.6.2 (Abstract Saddle Point Variational Problem) Let $a : U \times U \rightarrow \mathbb{R}$ and $b : U \times P \rightarrow \mathbb{R}$ be bounded bilinear operators and $l \in U'$ and $r \in P'$ be given. Find $(\mathbf{u}, p) \in U \times P$ such that

$$a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) = \langle l, \mathbf{v} \rangle,$$

$$b(\mathbf{u}, q) = \langle r, q \rangle,$$

with respect to an arbitrary $(\mathbf{v}, q) \in U \times P$. □

5.6.1 Existence Theorem of Solution

In reference to the existence of a unique solution to the abstract saddle point variational problem (Problem 5.6.2), the following result is known (cf. [55, Corollary 4.1, p. 61], [33, Theorem 1.1, p. 42], [95, Theorem 7.3, p. 135], [158, Theorem 4.3, p. 116]).

Theorem 5.6.3 (Solution to Abstract Saddle Point Variational Problem) Suppose $a : U \times U \rightarrow \mathbb{R}$ is a coercive and bounded bilinear operator on U_{div} (i.e., there exists some $\alpha > 0$ and

$$|a(\mathbf{v}, \mathbf{v})| \geq \alpha \|\mathbf{v}\|_U^2$$

is satisfied with respect to an arbitrary $\mathbf{v} \in U_{\text{div}}$). Also, let $b : U \times P \rightarrow \mathbb{R}$ be a bounded bilinear operator and that some $\beta > 0$ exists satisfying the inequality

$$\inf_{q \in P \setminus \{\mathbf{0}_P\}} \sup_{\mathbf{v} \in U \setminus \{\mathbf{0}_U\}} \frac{b(\mathbf{v}, q)}{\|\mathbf{v}\|_U \|q\|_P} \geq \beta. \quad (5.6.1)$$

In this case, the solution $(\mathbf{u}, p) \in U \times P$ to Problem 5.6.2 exists uniquely and with respect to $c > 0$ depending on α , β , $\|a\|$ and $\|b\|$,

$$\|\mathbf{u}\|_U + \|p\|_P \leq c (\|l\|_{U'} + \|r\|_{P'})$$

holds. □

Equation (5.6.1) is called the inf-sup condition or Ladyzenskaja–Babuška–Brezzi condition, Babuška–Brezzi–Kikuchi condition, etc.

If Theorem 5.6.3 is used, the existence of a unique solution to the Stokes problem can be shown in the following way.

Exercise 5.6.4 (Existence of Unique Solution to the Stokes Problem)

In Problem 5.5.2, it is supposed that some function $\tilde{\mathbf{u}} = \mathbf{u} - \mathbf{u}_D \in U_{\text{div}}$ exists and satisfies $\tilde{\mathbf{u}} = \mathbf{0}_{\mathbb{R}^d}$ on $\partial\Omega$. In this case, show that $(\tilde{\mathbf{u}}, p) \in U \times P$ satisfying Eqs. (5.5.14) and (5.5.15) exists uniquely. □

Answer Let us confirm that the assumptions of Theorem 5.6.3 hold in the following way. Let U and P be Hilbert spaces. Moreover, Problem 5.5.2 is equivalent to a problem seeking $(\tilde{\mathbf{u}}, p) \in U \times P$ which satisfies

$$a(\tilde{\mathbf{u}}, \mathbf{v}) + b(\mathbf{v}, p) = \hat{l}(\mathbf{v}), \quad b(\tilde{\mathbf{u}}, q) = \hat{r}(q),$$

with respect to an arbitrary $(\mathbf{v}, q) \in U \times P$, where

$$\hat{l}(\mathbf{v}) = l(\mathbf{v}) - a(\mathbf{u}_D, \mathbf{v}), \quad \hat{r}(q) = -b(\mathbf{u}_D, q).$$

We show that a is bounded and coercive on U_{div} . Clearly, a is bounded and coercive on U in view of Exercise 5.4.4. Next, we note that b is bounded and satisfies the inf-sup condition. In fact, when U_{div}^\perp is taken to be the orthogonal complement of U_{div} , an operator such that the domain of operator div is limited to U_{div}^\perp is taken to be τ . τ is bounded ($|\text{div} \mathbf{v}| / \|\mathbf{v}\|_U < \infty$), linear and injective (with respect to $\mathbf{v}_1, \mathbf{v}_2 \in U_{\text{div}}^\perp$, if $\tau \mathbf{v}_1 = \tau \mathbf{v}_2$, then $\mathbf{v}_1 = \mathbf{v}_2$). This is because with respect to $\mathbf{v} \in U_{\text{div}}^\perp$, if $\tau \mathbf{v} = \text{div} \mathbf{v} = 0$, then $\mathbf{v} \in U_{\text{div}}$, and we get $\mathbf{v} \in U_{\text{div}}^\perp \cap U_{\text{div}} = \{\mathbf{0}_U\}$. Furthermore,

it can be shown that τ is a surjection from U_{div}^\perp to P (see, e.g., [55] for the proof). Hence, the following inequalities,

$$\begin{aligned}
& \inf_{q \in P \setminus \{0_P\}} \sup_{\mathbf{v} \in U \setminus \{\mathbf{0}_U\}} \frac{b(\mathbf{v}, q)}{\|\mathbf{v}\|_U \|q\|_P} \\
&= \inf_{q \in P \setminus \{0_P\}} \sup_{\mathbf{v} \in U \setminus \{\mathbf{0}_U\}} \frac{(-\text{div} \mathbf{v}, q)_{L^2(\Omega; \mathbb{R})}}{\|\mathbf{v}\|_U \|q\|_P} \\
&\geq \inf_{q \in P \setminus \{0_P\}} \sup_{\mathbf{v} \in U \setminus \{\mathbf{0}_U\}} \frac{(-\tau \mathbf{v}, q)_P}{\|\tau^{-1}(-q)\|_U \|q\|_P} \geq \inf_{q \in P \setminus \{0_P\}} \frac{(q, q)_P}{\|\tau^{-1}(-q)\|_U \|q\|_P} \\
&\geq \frac{1}{\|\tau^{-1}\|_{\mathcal{L}(P; U_{\text{div}}^\perp)}} > 0,
\end{aligned}$$

are established. On the other hand, $\hat{l} \in U'$ is already verified in Exercise 5.4.4. Moreover, from assumptions of Eq. (5.5.1), we see that $\hat{r}(q) = 0 \in P'$. As seen above, Theorem 5.6.3 can be applied and $(\mathbf{u} - \mathbf{u}_D, p) \in U \times P$ satisfying Problem 5.5.2 exists uniquely. \square

5.6.2 Abstract Saddle Point Problem

In the abstract saddle point variational problem (Problem 5.6.2), if $a : U \times U \rightarrow \mathbb{R}$ is symmetric, its abstract saddle point variational problem is equivalent to an abstract saddle point problem such as the one below.

Problem 5.6.5 (Abstract Saddle Point Problem) Let $a : U \times U \rightarrow \mathbb{R}$ and $b : U \times P \rightarrow \mathbb{R}$ be bounded bilinear operators. Given $(l, r) \in U' \times P'$, define

$$\mathcal{L}(\mathbf{v}, q) = \frac{1}{2}a(\mathbf{v}, \mathbf{v}) + b(\mathbf{v}, q) - \langle l, \mathbf{v} \rangle - \langle r, q \rangle.$$

In this case, find $(\mathbf{u}, p) \in U \times P$ such that

$$\mathcal{L}(\mathbf{u}, q) \leq \mathcal{L}(\mathbf{u}, p) \leq \mathcal{L}(\mathbf{v}, p)$$

for any $(\mathbf{v}, q) \in U \times P$. \square

The following results can be obtained with respect to Problem 5.6.5 (cf. [55, Theorem 4.2, p. 62], [158, Theorem 4.4, p. 118]).

Theorem 5.6.6 (Agreement of Abstract Saddle Point Problems) *If a is symmetric ($a(\mathbf{u}, \mathbf{v}) = a(\mathbf{v}, \mathbf{u})$) and semi-positive definite (with respect to an arbitrary $\mathbf{v} \in U$, $a(\mathbf{v}, \mathbf{v}) \geq 0$ holds), then the solution to Problem 5.6.2 and the solution to Problem 5.6.5 agree.* \square

Let us check that in an abstract saddle point problem (Problem 5.6.5), $q \in P$ is a Lagrange multiplier with respect to equality constraint such as the continuity equation. Problem 5.6.5 is the problem, when setting

$$f(\mathbf{v}) = \frac{1}{2}a(\mathbf{v}, \mathbf{v}) - \langle l, \mathbf{v} \rangle,$$

seeking (\mathbf{v}, q) which satisfies

$$\min_{(\mathbf{v}, q) \in U \times P} \{ f(\mathbf{v}) \mid b(\mathbf{v}, q) - \langle r, q \rangle = 0 \}. \quad (5.6.2)$$

Here, $\mathcal{L}(\mathbf{v}, q)$ of Problem 5.6.5 is the Lagrange function of this problem and $q \in P$ is a Lagrange multiplier with respect to equality constraints. Theorem 5.6.6 which shows that the solution to Eq. (5.6.2) matches the saddle points of Problem 5.6.5 is a result corresponding to the duality theorem (Theorem 2.9.2).

5.7 Summary

In Chap. 5, we defined boundary value problems of elliptic partial differential equations, sought their weak form and studied the existence of a solution and its regularity. The key points from this chapter are as follows.

- (1) The existence of a unique solution for a boundary value problem of an elliptic partial differential equation (Poisson problem) is guaranteed when the assumptions of the Lax–Milgram theorem are satisfied with respect to the weak form (Sects. 5.1 and 5.2).
- (2) The regularity of a solution with respect to a boundary value problem of an elliptic partial differential equation depends on the regularity of given functions and the regularity of the boundary (Sect. 5.3).
- (3) A linear elastic problem is a boundary value problem of an elliptic partial differential equation. The existence of a unique solution with respect to this problem is guaranteed when the assumptions of the Lax–Milgram theorem are satisfied with respect to the weak form (Sect. 5.4).
- (4) A Stokes problem is a boundary value problem of an elliptic partial differential equation with a continuity equation as equality constraint. The existence of a unique solution of a Stokes problem is guaranteed when the assumptions for the existence of solutions with respect to an abstract saddle point variational problem using an inf-sup condition with respect to the weak form (Sect. 5.5) are satisfied.

5.8 Practice Problems

5.1 With respect to $b : \Omega \rightarrow \mathbb{R}$, $u_D : \Omega \rightarrow \mathbb{R}$, obtain the weak form of the boundary value problem seeking $u : \Omega \rightarrow \mathbb{R}$ satisfying

$$\begin{aligned} -\Delta u + u &= b && \text{in } \Omega, \\ u &= u_D && \text{on } \partial\Omega. \end{aligned}$$

Moreover, determine the appropriate function spaces for b and u_D with respect to the unique solution u of the corresponding weak form of the above system.

5.2 Consider a cantilever problem of a linear elastic body such as that in Fig. 5.12. In this case, show that, even though the point x_A is not a singular point, x_B is.

5.3 A dynamic linear elastic problem can be expressed as follows: “With respect to $b : \Omega \times (0, t_T) \rightarrow \mathbb{R}^d$, $p_N : \Gamma_N \times (0, t_T) \rightarrow \mathbb{R}^d$, $u_D : \Omega \times (0, t_T) \rightarrow \mathbb{R}^d$, $u_{D0} : \Omega \rightarrow \mathbb{R}^d$, $u_{DT} : \Omega \rightarrow \mathbb{R}^d$ and $\rho > 0$, obtain $u : \Omega \times (0, t_T) \rightarrow \mathbb{R}^d$ which satisfies

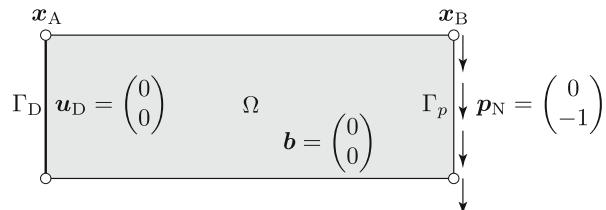
$$\begin{aligned} \rho \ddot{u}^\top - \nabla \cdot S(u) &= b^\top && \text{in } \Omega \times (0, t_T), \\ S(u) v &= p_N && \text{on } \Gamma_N \times (0, t_T), \\ u &= u_D && \text{on } \Gamma_D \times (0, t_T), \\ u &= u_{D0} && \text{in } \Omega \times \{0\}, \\ u &= u_{DT} && \text{in } \Omega \times \{t_T\}, \end{aligned}$$

where $\dot{u} = \partial u / \partial t$ with respect to time $t \in (0, t_T)$.” Obtain the weak form of this problem.

5.4 In Exercise 5.3, when $b = \mathbf{0}_{\mathbb{R}^d}$, $p_N = \mathbf{0}_{\mathbb{R}^d}$ and $u_D = \mathbf{0}_{\mathbb{R}^d}$, when with respect to $(x, t) \in \Omega \times (0, t_T)$, a solution of variable separation (standing wave) is assumed to be

$$u(x, t) = \phi(x) e^{\lambda t}.$$

Fig. 5.12 Cantilever problem of linear elastic body



The problem of seeking $\phi : \Omega \rightarrow \mathbb{R}^d$ and $\lambda \in \mathbb{R}$ is called an eigenfrequency problem. Obtain the weak form of this problem.

5.5 With respect to $\mathbf{b} : \Omega \times (0, t_T) \rightarrow \mathbb{R}^d$, $\mathbf{u}_D : \Omega \times (0, t_T) \rightarrow \mathbb{R}^d$, $\mu > 0$ and $\rho > 0$, the problem seeking $(\mathbf{u}, p) : \Omega \times (0, t_T) \rightarrow \mathbb{R}^d \times \mathbb{R}$ which satisfies

$$\begin{aligned} \rho \dot{\mathbf{u}} + \rho (\mathbf{u} \cdot \nabla) \mathbf{u} - \mu \Delta \mathbf{u} + \nabla p &= \mathbf{b} \quad \text{in } \Omega \times (0, t_T), \\ \nabla \cdot \mathbf{u} &= 0 \quad \text{in } \Omega \times (0, t_T), \\ \mathbf{u} &= \mathbf{u}_D \text{ on } \{\partial\Omega \times (0, t_T)\} \cup \{\Omega \times \{0\}\} \end{aligned}$$

is called a Navier–Stokes problem. The first equation is called a Navier–Stokes equation and the second equation is called a continuity equation. Obtain the weak form of this problem.

5.6 With respect to an isotropic linear elastic body, show that the relationship $e_Y = 2\mu_L(1 + \nu_P)$ holds between Young's modulus e_Y , elastic shear modulus μ_L and Poisson's ratio ν_P .

Chapter 6

Fundamentals of Numerical Analysis



In Chap. 5, covering several boundary value problems of elliptic partial differential equations, we saw that the existence of their unique solutions could be guaranteed using solutions of the weak form. These will be referred to as the exact solutions. Exact solutions can be found analytically if the shape of the domain is somewhat simple such as a rectangle or an ellipse. However, difficulties arise for domains whose shape may have arbitrarily moved as examined in this book. In order to solve a shape optimization problem, even if an exact solution is not possible, one can resort to a numerical analysis method to obtain an approximate solution.

This chapter looks at how to obtain an approximate solution to such boundary value problems. The Galerkin method is considered first in this chapter. In the Galerkin method, approximate functions with respect to the solution function and an arbitrarily selected variational function (test function) are constructed with linear combinations of given basis functions multiplied by undetermined multipliers which are determined by substituting the approximate functions into the weak form. The characteristics of this method are not only its clarity but also the fact that the unique existence of the approximate solution is guaranteed by the Lax–Milgram theorem shown in Chap. 5. However, constraints due to the shape of the domain may come into play depending on the choice of basis function.

In order to remove such constraints, the finite element method is introduced to devise an appropriate selection of the basis function. Furthermore, error analysis is possible for approximate solutions using the finite element method. These results show that the finite element method is flexible for domain shape approximation and provides a good method for reliably guaranteeing the convergence to an exact solution if the divisions into elements are increased.

6.1 Galerkin Method

Here, we examine how the Galerkin method can be applied to approximate a solution for a boundary problems of elliptic partial differential equation. This will be illustrated by solving a one-dimensional Poisson problem and a $d \in \{2, 3\}$ -dimensional Poisson problem.

6.1.1 One-Dimensional Poisson Problem

Consider the following one-dimensional Poisson problem with a mixed boundary condition.

Problem 6.1.1 (One-Dimensional Poisson Problem) Let the functions $b : (0, 1) \rightarrow \mathbb{R}$, $p_N \in \mathbb{R}$ and $u_D : (0, 1) \rightarrow \mathbb{R}$ be given. Find $u : (0, 1) \rightarrow \mathbb{R}$ such that

$$-\frac{d^2u}{dx^2} = b \quad \text{in } (0, 1), \quad \frac{du}{dx}(1) = p_N, \quad u(0) = u_D(0). \quad \square$$

With respect to Problem 6.1.1, set

$$U = \left\{ v \in H^1((0, 1); \mathbb{R}) \mid v(0) = 0 \right\}. \quad (6.1.1)$$

Moreover, with respect to $u, v \in U$, let

$$a(u, v) = \int_0^1 \frac{du}{dx} \frac{dv}{dx} dx, \quad (6.1.2)$$

$$l(v) = \int_0^1 bv dx + p_N v(1). \quad (6.1.3)$$

The weak form of this problem is as follows.

Problem 6.1.2 (Weak Form of 1D Poisson Problem) Let $a(\cdot, \cdot)$ and $l(\cdot)$ be given by Eqs. (6.1.2) and (6.1.3), respectively. For given functions $b \in L^2((0, 1); \mathbb{R})$, $p_N \in \mathbb{R}$ and $u_D \in H^1((0, 1); \mathbb{R})$, obtain $u - u_D \in U$ satisfying

$$a(u, v) = l(v) \quad (6.1.4)$$

with respect to an arbitrary $v \in U$. \square

In the Galerkin method, approximate functions are constructed in the following way with respect to Problem 6.1.2.

Definition 6.1.3 (Set of Approximate Functions) Let $\phi = (\phi_1, \dots, \phi_m)^\top \in U^m$ be $m \in \mathbb{N}$ known linearly independent functions. Let the set of approximate functions for U be

$$U_h = \left\{ v_h(\alpha) = \sum_{i \in \{1, \dots, m\}} \alpha_i \phi_i = \alpha \cdot \phi \mid \alpha \in \mathbb{R}^m \right\}$$

with $\alpha = (\alpha_i)_i \in \mathbb{R}^m$ as undetermined multipliers. In this case, ϕ is called a basis function. \square

Figure 6.1 shows an example of u_D and ϕ . By using u_D and U_h defined in this way, the approximate functions of $u - u_D \in U$ and $v \in U$ can be supposed to be $u_h - u_D \in U_h$ and $v_h \in U_h$. If using Definition 4.2.6, this can be written as

$$u_h = u_D + \text{span } \phi,$$

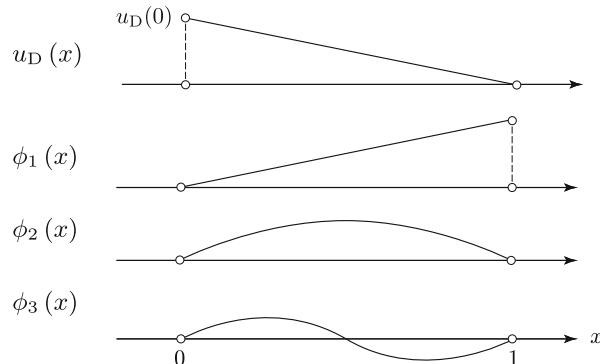
$$v_h = \text{span } \phi.$$

In this case, U_h becomes a linear subspace (linear space) containing ϕ . Furthermore, if the inner product defined with respect to $H^1((0, 1); \mathbb{R})$ is used, U_h becomes a Hilbert space. Here, the number of vectors which can be independently selected within U_h is m . In this case, U_h is a Hilbert space with m dimensions.

We have just defined the set of approximate functions U_h and u_D . Hence, using these functions we can now define the Galerkin method with respect to Problem 6.1.2 in the following way.

Definition 6.1.4 (Galerkin Method) Let u_D and U_h be as in Definition 6.1.3. If $u_h(\alpha) - u_D \in U_h$ and $v_h(\beta) \in U_h$ are substituted into $u - u_D \in U$ and $v \in U$ of Eq. (6.1.4), simultaneous linear equations with $\alpha \in \mathbb{R}^m$ as an unknown vector can be obtained. Using the solution α to these equations, the method for obtaining the approximate solution of Problem 6.1.2 using $u_h(\alpha) = u_D + \phi \cdot \alpha$ is known as the Galerkin method. \square

Fig. 6.1 An example of u_D and basis functions ϕ



From the fact that U_h is a Hilbert space, if the Lax–Milgram theorem is applied, the solution by the Galerkin method $u_h(\alpha)$ can be said to exist uniquely. Actually, it is because $a(\cdot, \cdot)$ is bounded, $a(\cdot, \cdot)$ is coercive on $U_h \times U_h$ since U_h satisfies the homogeneous boundary condition, and $\hat{l}(\cdot)$ such as Exercise 5.2.5 is included in U'_h . Furthermore, $a(\cdot, \cdot)$ is also symmetric. The symmetry and coerciveness of $a(\cdot, \cdot)$ will appear later as the symmetry and positive definitiveness of coefficient matrix in simultaneous linear equations.

Let us look at how we can actually solve Problem 6.1.2 using the Galerkin method. Use u_D and U_h of Definition 6.1.3 to substitute $u_h - u_D \in U_h$ and $v_h \in U_h$ into the weak form (Eq. (6.1.4)) and seek to find u_h such that

$$a(u_h, v_h) = l(v_h) \quad (6.1.5)$$

is satisfied with respect to an arbitrary $v_h \in U_h$. Both sides of Eq. (6.1.5) are respectively given by

$$\begin{aligned} a(u_h, v_h) &= \int_0^1 \frac{du_h}{dx} \frac{dv_h}{dx} dx, \\ l(v_h) &= \int_0^1 bv_h dx + p_N v_h(1). \end{aligned}$$

The terms in the integrand of $a(u_h, v_h)$ are given by

$$\begin{aligned} \frac{du_h}{dx} &= \frac{du_D}{dx} + \frac{d\phi}{dx} \cdot \alpha = \frac{du_D}{dx} + \left(\frac{d\phi_1}{dx} \dots \frac{d\phi_m}{dx} \right) \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix}, \\ \frac{dv_h}{dx} &= \frac{d\phi}{dx} \cdot \beta = \left(\frac{d\phi_1}{dx} \dots \frac{d\phi_m}{dx} \right) \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}. \end{aligned}$$

Hence, $a(u_h, v_h)$ becomes

$$\begin{aligned} a(u_h, v_h) &= \int_0^1 \frac{du_h}{dx} \frac{dv_h}{dx} dx \\ &= \int_0^1 (\beta_1 \dots \beta_m) \begin{pmatrix} \frac{d\phi_1}{dx} \\ \vdots \\ \frac{d\phi_m}{dx} \end{pmatrix} \left(\frac{du_D}{dx} + \left(\frac{d\phi_1}{dx} \dots \frac{d\phi_m}{dx} \right) \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} \right) dx \end{aligned}$$

$$\begin{aligned}
&= \int_0^1 (\beta_1 \cdots \beta_m) \\
&\quad \times \left(\begin{pmatrix} \frac{d\phi_1}{dx} \frac{d\phi_1}{dx} & \cdots & \frac{d\phi_1}{dx} \frac{d\phi_m}{dx} \\ \vdots & \ddots & \vdots \\ \frac{d\phi_m}{dx} \frac{d\phi_1}{dx} & \cdots & \frac{d\phi_m}{dx} \frac{d\phi_m}{dx} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} + \begin{pmatrix} \frac{du_D}{dx} \frac{d\phi_1}{dx} \\ \vdots \\ \frac{du_D}{dx} \frac{d\phi_m}{dx} \end{pmatrix} \right) dx \\
&= (\beta_1 \cdots \beta_m) \\
&\quad \times \left(\begin{pmatrix} a(\phi_1, \phi_1) & \cdots & a(\phi_1, \phi_m) \\ \vdots & \ddots & \vdots \\ a(\phi_m, \phi_1) & \cdots & a(\phi_m, \phi_m) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} + \begin{pmatrix} a(u_D, \phi_1) \\ \vdots \\ a(u_D, \phi_m) \end{pmatrix} \right) \\
&= \boldsymbol{\beta} \cdot (\boldsymbol{A}\boldsymbol{\alpha} + \boldsymbol{a}_D).
\end{aligned}$$

Here, we wrote $\boldsymbol{A} = (a_{ij})_{(i,j) \in \{1, \dots, m\}^2}$ and $\boldsymbol{a}_D = (a_{Di})_{i \in \{1, \dots, m\}}$, and let

$$a_{ij} = a(\phi_i, \phi_j) = \int_0^1 \frac{d\phi_i}{dx} \frac{d\phi_j}{dx} dx, \quad (6.1.6)$$

$$a_{Di} = a(u_D, \phi_i) = \int_0^1 \frac{du_D}{dx} \frac{d\phi_i}{dx} dx. \quad (6.1.7)$$

One the other hand, $\boldsymbol{l}(v_h)$ becomes

$$\begin{aligned}
\boldsymbol{l}(v_h) &= \int_0^1 b v_h dx + p_N v_h(1) \\
&= \int_0^1 b (\beta_1 \cdots \beta_m) \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_m \end{pmatrix} dx + p_N (\beta_1 \cdots \beta_m) \begin{pmatrix} \phi_1(1) \\ \vdots \\ \phi_m(1) \end{pmatrix} \\
&= (\beta_1 \cdots \beta_m) \begin{pmatrix} l(\phi_1) \\ \vdots \\ l(\phi_m) \end{pmatrix} = \boldsymbol{\beta} \cdot \boldsymbol{l}.
\end{aligned}$$

Here, we let $\boldsymbol{l} = (l_i)_{i \in \{1, \dots, m\}}$ and

$$l_i = l(\phi_i) = \int_0^1 b \phi_i dx + p_N \phi_i(1). \quad (6.1.8)$$

Therefore, Eq. (6.1.5) can be written as

$$\boldsymbol{\beta} \cdot (\mathbf{A}\boldsymbol{\alpha} + \mathbf{a}_D) = \boldsymbol{\beta} \cdot \mathbf{l}.$$

Here, consider an arbitrary $\boldsymbol{\beta} \in \mathbb{R}^m$ to get

$$\mathbf{A}\boldsymbol{\alpha} = \mathbf{l} - \mathbf{a}_D = \hat{\mathbf{l}}. \quad (6.1.9)$$

For confirmation, write the elements of vector and matrix of Eq. (6.1.9) to get

$$\begin{pmatrix} a(\phi_1, \phi_1) & \cdots & a(\phi_1, \phi_m) \\ \vdots & \ddots & \vdots \\ a(\phi_m, \phi_1) & \cdots & a(\phi_m, \phi_m) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} = \begin{pmatrix} l(\phi_1) - a(u_D, \phi_1) \\ \vdots \\ l(\phi_m) - a(u_D, \phi_m) \end{pmatrix}.$$

\mathbf{A} is referred to as a coefficient matrix and $\hat{\mathbf{l}}$ is a known term vector. \mathbf{A} is symmetric from $a(\phi_i, \phi_j) = a(\phi_j, \phi_i)$. Moreover, due to the coerciveness of $a(\cdot, \cdot)$ at U_h , there exists some $c_0 > 0$ and

$$\boldsymbol{\beta} \cdot (\mathbf{A}\boldsymbol{\beta}) = a(\mathbf{v}_h, \mathbf{v}_h) \geq c_0 \|\boldsymbol{\beta}\|_{\mathbb{R}^m}^2$$

holds with respect to an arbitrary $\boldsymbol{\beta} \in \mathbb{R}^m$. Therefore, \mathbf{A} becomes a positive definite symmetric matrix (hence a regular matrix) and the inverse matrix of \mathbf{A} can be taken to calculate $\boldsymbol{\alpha}$ by

$$\boldsymbol{\alpha} = \mathbf{A}^{-1}\hat{\mathbf{l}}. \quad (6.1.10)$$

The approximate solution u_h can be obtained via

$$u_h = u_D + \boldsymbol{\phi} \cdot \boldsymbol{\alpha} \quad (6.1.11)$$

by using this $\boldsymbol{\alpha}$.

Based on the detail above, basis function u_D and $\boldsymbol{\phi}$ should be selected so that the calculation of $\int_0^1 (d\phi_i/dx)(d\phi_j/dx) dx$ is easily possible. Furthermore, if selection can be made so that the derivatives of the basis function are mutually orthogonal, \mathbf{A} becomes a diagonal matrix and the calculation of the inverse matrix becomes simple.

We solve the following problem using the Galerkin method.

Exercise 6.1.5 (Galerkin Method for 1D Dirichlet Problem) Let the basis functions be

$$\boldsymbol{\phi} = (\phi_1 \cdots \phi_m)^\top = (\sin(1\pi x) \cdots \sin(m\pi x))^\top$$

and use the Galerkin method in order to obtain the approximate solution $u_h : (0, 1) \rightarrow \mathbb{R}$ satisfying

$$-\frac{d^2u}{dx^2} = 1 \quad \text{in } (0, 1), \quad u(0) = 0, \quad u(1) = 0. \quad \square$$

Answer Let $U = H_0^1((0, 1); \mathbb{R})$ and write the weak form of this problem as

$$a(u, v) = l_1(v)$$

with respect to an arbitrary $v \in U$. Here, let $a(\cdot, \cdot)$ be Eq. (6.1.2) and $l_1(\cdot)$ be $l(\cdot)$ on Eq. (6.1.3) when $b = 1$ and $p = 0$. Let the approximate function with respect to $\alpha, \beta \in \mathbb{R}^m$ be

$$u_h = \alpha \cdot \phi(x), \quad v_h = \beta \cdot \phi(x).$$

Substituting these into the weak form gives

$$\beta \cdot (A\alpha) = \beta \cdot l_1.$$

Here, let $A = (a_{ij})_{(i,j) \in \{1, \dots, m\}^2}$ and $l_1 = (l_{1i})_{i \in \{1, \dots, m\}}$ to get

$$\begin{aligned} a_{ij} &= a(\phi_i, \phi_j) = \int_0^1 \frac{d\phi_i}{dx} \frac{d\phi_j}{dx} dx = ij\pi^2 \int_0^1 \cos(i\pi x) \cos(j\pi x) dx \\ &= \frac{1}{2}ij\pi^2 \int_0^1 [\cos((i+j)\pi x) + \cos((i-j)\pi x)] dx = \frac{1}{2}ij\pi^2 \delta_{ij}, \\ l_{1i} &= \int_0^1 \sin(i\pi x) dx = \frac{1}{i\pi} [-\cos(i\pi x)]_0^1 = \frac{(-1)^{i+1} + 1}{i\pi}, \end{aligned}$$

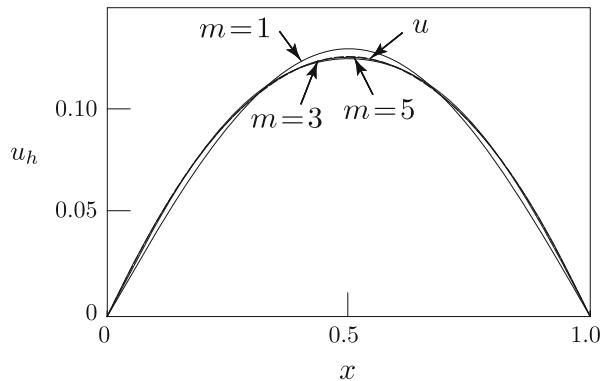
where

$$\delta_{ij} = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j) \end{cases}$$

represents the Kronecker delta. Hence, $A\alpha = l_1$ becomes

$$\frac{\pi^2}{2} \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 4 & 0 & \cdots & 0 \\ 0 & 0 & 9 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & m^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_m \end{pmatrix} = \frac{1}{\pi} \begin{pmatrix} 2 \\ 0 \\ 2/3 \\ \vdots \\ \frac{(-1)^{m+1} + 1}{m} \end{pmatrix}.$$

Fig. 6.2 Exact and approximate solutions of Exercise 6.1.5



Solving these simultaneous linear equations gives

$$\alpha_i = \frac{2 \{(-1)^{i+1} + 1\}}{i^3 \pi^3}.$$

Therefore, an approximate solution becomes

$$u_h = \sum_{i \in \{1, \dots, m\}} \frac{2 \{(-1)^{i+1} + 1\}}{i^3 \pi^3} \sin(i \pi x).$$

On the other hand, the exact solution is

$$u = \frac{1}{2}x(x-1).$$

Figure 6.2 shows the comparison between the approximate solution and the exact solution. \square

6.1.2 *d*-Dimensional Poisson Problem

Next, to think about solving the $d \in \{2, 3\}$ -dimensional Poisson problem using the Galerkin method, we revisit Problem 5.1.1.

Problem 6.1.6 (*d*-Dimensional Poisson Problem) When $b : \Omega \rightarrow \mathbb{R}$, $p_N : \Gamma_N \rightarrow \mathbb{R}$ and $u_D : \Omega \rightarrow \mathbb{R}$ are given, obtain $u : \Omega \rightarrow \mathbb{R}$ which satisfies

$$-\Delta u = b \quad \text{in } \Omega, \quad \frac{\partial u}{\partial \nu} = p_N \quad \text{on } \Gamma_N, \quad u = u_D \quad \text{on } \Gamma_D. \quad \square$$

With respect to Problem 6.1.6, let

$$U = \left\{ u \in H^1(\Omega; \mathbb{R}) \mid u = 0 \text{ on } \Gamma_D \right\}. \quad (6.1.12)$$

Moreover, with respect to $u, v \in U$, let

$$a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v \, dx, \quad (6.1.13)$$

$$l(v) = \int_{\Omega} bv \, dx + \int_{\Gamma_N} p_N v \, d\gamma. \quad (6.1.14)$$

The weak form of this problem becomes as follows.

Problem 6.1.7 (Weak Form of d -Dimensional Poisson Problem) Given $b \in L^2(\Omega; \mathbb{R})$, $p_N \in L^2(\Gamma_N; \mathbb{R})$ and $u_D \in H^1(\Omega; \mathbb{R})$, obtain $u - u_D \in U$ such that

$$a(u, v) = l(v) \quad (6.1.15)$$

is satisfied with respect to an arbitrary $v \in U$. \square

Construct approximate functions with respect to Problem 6.1.6 in the following way. Let m be a natural number.

Definition 6.1.8 (Set of Approximate Functions) Let $\phi = (\phi_1, \dots, \phi_m)^\top \in U^m$ be m known functions of linear independence. Let

$$U_h = \left\{ v_h(\alpha) = \sum_{i \in \{1, \dots, m\}} \alpha_i \phi_i = \alpha \cdot \phi \mid \alpha \in \mathbb{R}^m \right\}$$

be the set of approximate functions with respect to U with $\alpha = (\alpha_i)_i \in \mathbb{R}^m$ as unknown multipliers. Here, ϕ is called basis functions. \square

Comparing Definition 6.1.3 and Definition 6.1.8 shows that the same expression is being used other than the defined domain of the function being changed. Hence, the same set of equations used for the one-dimensional Poisson problem can be used for the d -dimensional Poisson problem. This is confirmed as follows. When $u_h - u_D \in U_h$ and $v_h \in U_h$ with u_D and U_h in Definition 6.1.8 are substituted into the weak form (Eq. (6.1.15)),

$$a(u_h, v_h) = l(v_h) \quad (6.1.16)$$

becomes the same as Eq. (6.1.5). Here, the definitions of $a(\cdot, \cdot, \cdot)$ and $l(\cdot)$ are replaced by

$$a(u_h, v_h) = \int_{\Omega} \nabla u_h \cdot \nabla v_h \, dx,$$

$$l(v_h) = \int_{\Omega} b v_h \, dx + \int_{\Gamma_N} p_N v_h \, d\gamma.$$

After this, the same expansion of equations used for the one-dimensional Poisson problem can be performed to obtain

$$A\alpha = l - a_D = \hat{l} \quad (6.1.17)$$

which coincides with Eq. (6.1.9), where $A = (a_{ij})_{(i,j) \in \{1, \dots, m\}^2}$, $a_D = (a_{Di})_{i \in \{1, \dots, m\}}$ and $l = (l_i)_{i \in \{1, \dots, m\}}$ are changed by

$$a_{ij} = a(\phi_i, \phi_j) = \int_{\Omega} \nabla \phi_i \cdot \nabla \phi_j \, dx, \quad (6.1.18)$$

$$a_{Di} = a(u_D, \phi_i) = \int_{\Omega} \nabla u_D \cdot \nabla \phi_i \, dx, \quad (6.1.19)$$

$$l_i = l(\phi_i) = \int_{\Omega} b \phi_i \, dx + \int_{\Gamma_N} p_N \phi_i \, d\gamma, \quad (6.1.20)$$

respectively. Since A becomes a positive definite symmetric matrix, and the inverse matrix of A can be taken, Eqs. (6.1.10) and (6.1.11) remain in effect.

We solve in the following exercise a two-dimensional Dirichlet problem using the Galerkin method with respect to a square domain. Looking at this it is apparent that the notations of Eqs. (6.1.10) and (6.1.11) are established [96, Exercise 3.2, p. 30].

Exercise 6.1.9 (Galerkin Method for 2D Dirichlet Problem) Setting basis functions as

$$\phi = (\phi_{ij}(x))_{(i,j) \in \{1, \dots, m\}^2} = (\sin(i\pi x_1) \sin(j\pi x_2))_{(i,j) \in \{1, \dots, m\}^2},$$

use the Galerkin method in order to obtain the approximate solution $u_h : (0, 1)^2 \rightarrow \mathbb{R}$ which satisfies

$$-\Delta u = 1 \quad \text{in } \Omega = (0, 1)^2, \quad u = 0 \quad \text{on } \partial\Omega. \quad \square$$

Answer Let $U = H_0^1((0, 1)^2; \mathbb{R})$ and denote the weak form of this problem as

$$a(u, v) = l_1(v)$$

with respect to an arbitrary $v \in U$, where we let $a(\cdot, \cdot)$ be as in Eq. (6.1.13) and $l_1(\cdot)$ be $l(\cdot)$ in Eq. (6.1.14) when $b = 1$ and $p = 0$. Let the approximate functions be

$$u_h = \boldsymbol{\alpha} \cdot \boldsymbol{\phi}(\mathbf{x}) = \sum_{(i,j) \in \{1, \dots, m\}^2} \alpha_{ij} \phi_{ij}(\mathbf{x}),$$

$$v_h = \boldsymbol{\beta} \cdot \boldsymbol{\phi}(\mathbf{x}) = \sum_{(i,j) \in \{1, \dots, m\}^2} \beta_{ij} \phi_{ij}(\mathbf{x})$$

with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta} \in \mathbb{R}^{m \times m}$. Substituting u_h and v_h into the weak form gives

$$\boldsymbol{\beta} \cdot (A\boldsymbol{\alpha}) = \boldsymbol{\beta} \cdot \mathbf{l}_1.$$

Here, if we write $A = (a(\phi_{ij}, \phi_{kl}))_{(i,j,k,l) \in \{1, \dots, m\}^4}$ and $\mathbf{l}_1 = (l_1(\phi_{ij}))_{(i,j) \in \{1, \dots, m\}^2}$,

$$\begin{aligned} a(\phi_{ij}, \phi_{kl}) &= \int_0^1 \int_0^1 \left(\frac{\partial \phi_{ij}}{\partial x_1} \frac{\partial \phi_{kl}}{\partial x_1} + \frac{\partial \phi_{ij}}{\partial x_2} \frac{\partial \phi_{kl}}{\partial x_2} \right) dx_1 dx_2 \\ &= \int_0^1 \int_0^1 \left\{ ki\pi^2 \cos(k\pi x_1) \sin(l\pi x_2) \cos(i\pi x_1) \sin(j\pi x_2) \right. \\ &\quad \left. + li\pi^2 \sin(k\pi x_1) \cos(l\pi x_2) \sin(i\pi x_1) \cos(j\pi x_2) \right\} dx_1 dx_2 \\ &= \frac{\pi^2}{4} (ki + lj) \delta_{ki} \delta_{lj}, \\ l_1(\phi_{ij}) &= \int_0^1 \int_0^1 \sin(i\pi x_1) \sin(j\pi x_2) dx_1 dx_2 \\ &= \frac{\{(-1)^{i+1} + 1\} \{(-1)^{j+1} + 1\}}{ij\pi^2} \end{aligned}$$

is obtained. Then $A\boldsymbol{\alpha} = \mathbf{l}_1$ becomes

$$\sum_{(k,l) \in \{1, \dots, m\}^2} \frac{\pi^2}{4} (ki + lj) \delta_{ki} \delta_{lj} \alpha_{ij} = \frac{\{(-1)^{i+1} + 1\} \{(-1)^{j+1} + 1\}}{ij\pi^2}.$$

Solving these simultaneous first-order equations gives

$$\alpha_{ij} = \frac{4 \{(-1)^{i+1} + 1\} \{(-1)^{j+1} + 1\}}{ij(i^2 + j^2)\pi^4}$$

with respect to $i, j \in \{1, \dots, m\}$. Therefore, the approximate solution u_h becomes

$$u_h = \sum_{(i,j) \in \{1, \dots, m\}^2} \frac{4 \{(-1)^{i+1} + 1\} \{(-1)^{j+1} + 1\}}{ij (i^2 + j^2) \pi^4} \sin(i\pi x) \sin(j\pi x). \quad (6.1.21)$$

□

6.1.3 Ritz Method

In the Galerkin method, an approximate function was substituted into the weak form. However, the same approximate solution can also be obtained when an approximate function is substituted into a minimization problem. This method is called the Ritz method.

Take a look at the next problem rewriting Problem 6.1.7 as a minimization problem. $a(\cdot, \cdot)$, $l(\cdot)$, u_D and U are taken to be the same as those used in Problem 6.1.7.

Problem 6.1.10 (Minimization Problem of dD Poisson Problem) With respect to $b \in L^2(\Omega; \mathbb{R})$, $p_N \in L^2(\Gamma_N; \mathbb{R})$ and $u_D \in H^1(\Omega; \mathbb{R})$, obtain a u which satisfies

$$\min_{u-u_D \in U} \left\{ f(u) = \frac{1}{2} a(u, u) - l(u) \right\}. \quad \square$$

The Ritz method is defined as follows with respect to Problem 6.1.10.

Definition 6.1.11 (Ritz Method) Let U_h be as in Definition 6.1.8. If $u_h(\alpha) - u_D \in U_h$ is substituted into $u - u_D \in U$ in Problem 6.1.10, the stationary condition of $f(u)$ with respect to a variation of $\alpha \in \mathbb{R}^m$ can be obtained as simultaneous linear equations with respect to α . The method for obtaining an approximate solution of Problem 6.1.10 using the solution α via $u_h(\alpha) = u_D + \phi \cdot \alpha$ is called the Ritz method. □

We solve Problem 6.1.10 using the Ritz method. Substituting u_h into $f(u_h)$ gives

$$f(u_h) = \frac{1}{2} \{a(u_D, u_D) + \alpha \cdot (A\alpha) + 2\alpha \cdot a_D\} - (l(u_D) + \alpha \cdot l),$$

where A , a_D and l are Eqs. (6.1.18), (6.1.19) and (6.1.20) respectively.

From the minimum conditions of $f(u_h)$,

$$\frac{\partial f(u_h)}{\partial \alpha} = A\alpha + a_D - l = \mathbf{0}_{\mathbb{R}^m}$$

is obtained. This equation matches Eq. (6.1.17) of the Galerkin method.

The fact that the Ritz method just requires the construction of an approximate function with respect to u means it has an advantage in a sense that the theory can be compactly contained. However, compared with the Galerkin method, it should be noted that its use is limited to when the boundary value problem of elliptic partial differential equation can be rewritten as a minimization problem, namely $a(\cdot, \cdot)$ is symmetric, as seen in Sect. 5.2. Moreover, when $a(\cdot, \cdot)$ is symmetric, the two methods can be put together because the equation obtained from the Galerkin method is the same as that of the Ritz method, and referred to as the Ritz–Galerkin method.

6.1.4 Basic Error Estimation

The existence of a unique solution $u_h(\alpha)$ by the Galerkin method was guaranteed by the Lax–Milgram theorem from the fact that the set of approximate functions U_h is a Hilbert space. Furthermore, if using the norm of the Hilbert space containing the exact solution, clear results can be obtained regarding the stability and errors of the approximate solution.

Here, with respect to Problem 6.1.1,

$$\begin{aligned} V &= H^1((0, 1); \mathbb{R}), \\ U &= \{v \in V \mid v(0) = 0\}, \\ U(u_D) &= \{v \in V \mid v - u_D \in U, u_D \in V\} \end{aligned}$$

is set. With respect to Problem 6.1.6, let

$$\begin{aligned} V &= H^1(\Omega; \mathbb{R}), \\ U &= \{v \in V \mid v = 0 \text{ on } \Gamma_D\}, \\ U(u_D) &= \{v \in V \mid v - u_D \in U, u_D \in V\}. \end{aligned}$$

Let V' be the dual space of V . Moreover, the set of approximate functions U_h is as per Definition 6.1.3 with respect to Problem 6.1.1 and Definition 6.1.8 with respect to Problem 6.1.6. Furthermore, the approximate function of u_D has not been thought about but here, its approximate function is set to be u_{Dh} and written as

$$U_h(u_{Dh}) = \{u_h \in V \mid h_h - u_{Dh} \in U_h\}.$$

One of the results is like the one below relating to the effect that the error in the given functions has on the approximate solution (cf. [42, Remark 1.2, p. 30], [95, Theorem 2.3, p. 34]).

Theorem 6.1.12 (Stability of Approximate Solution) *Let $a : V \times V \rightarrow \mathbb{R}$ be a bounded and coercive bilinear form. Let $u_{h1}, u_{h2} \in U_h (u_{Dh}) \subset U (u_D)$ be the approximate solutions by the Galerkin method of Problem 6.1.2 or Problem 6.1.7 with respect to arbitrary $l_1, l_2 \in V'$ respectively. In this case,*

$$\|u_{h1} - u_{h2}\|_V \leq \frac{1}{\alpha} \|l_1 - l_2\|_{V'}$$

is established. Here, $\alpha > 0$ is a constant representing the coerciveness of $a(\cdot, \cdot)$ (Definition 5.2.1). \square

The other result shows that the approximate solution of the Galerkin method is the best (closest to the exact solution) element among the set of approximate solutions $U_h (u_{Dh})$. Here, the distance between the exact solution $u \in U (u_D)$ and $u_h \in U_h (u_{Dh})$ will be measured with $\sqrt{a(u - u_h, u - u_h)}$ or the norm $\|u - u_h\|_V$ in V . The result is called Cea's lemma (cf. [42, Theorem 13.1, p. 113], [157, Lemma 2.3, p. 54], [95, Theorem 2.4, p. 42]).

Theorem 6.1.13 (Basic Error Estimation) *Let $u \in U (u_D)$ be the solution to Problem 6.1.2 or Problem 6.1.7 with respect to an arbitrary $l \in V'$, and $u_h \in U_h (u_{Dh})$ be the approximate solution from the Galerkin method. In this case,*

$$\begin{aligned} a(u - u_h, u - u_h) &\leq \inf_{v_h \in U_h (u_{Dh})} a(u - v_h, u - v_h), \\ \|u - u_h\|_V &\leq \sqrt{\frac{\|a\|}{\alpha}} \inf_{v_h \in U_h (u_{Dh})} \|u - v_h\|_V + \left(1 + \sqrt{\frac{\|a\|}{\alpha}}\right) \|u_D - u_{Dh}\|_V \end{aligned}$$

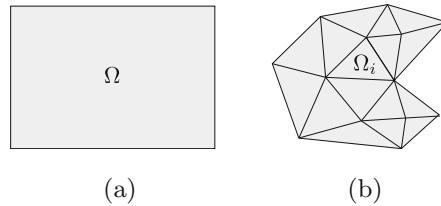
is established. Here, $\|a\|$ is the norm of bilinear operator (Sect. 4.4.4) and $\alpha > 0$ is a constant providing the coerciveness of $a(\cdot, \cdot)$. \square

Theorem 6.1.13 shows that the approximate solution from the Galerkin method is the best one out of U_h . Hence, it indicates that in order to reduce the error of approximate solution from the Galerkin method, it is effective to keep the approximate functions with the ability to be close to the exact solution within U_h . This result is also used when conducting error estimation of numerical analyses based on the Galerkin method. The error estimation with respect to finite element methods is examined in detail in Sect. 6.6.

6.2 One-Dimensional Finite Element Method

An approximate function is constructed as a linear combination of basis functions in the Galerkin method and the undetermined multipliers are found by substituting them into the weak form. Let us consider how to choose the basis function without changing this framework.

Fig. 6.3 Supports of approximate functions. (a) Ω : Galerkin method in Sect. 6.1. (b) $\{\Omega_i\}_i$: Finite element method



In the Galerkin method seen in Sect. 6.1, the basis functions were selected from the functions defined across the whole domain. In Exercise 6.1.9, as the basis functions, $(\sin(i\pi x_1) \sin(j\pi x_2))_{(i,j) \in \{1, \dots, m\}^2}$ were chosen. These functions have the support on the domain $\Omega = (0, 1)^2$ of the boundary value problem of partial differential equation. As long as the basis functions are chosen in this way, the shape of the domain on which the boundary value problem is defined will be limited to be a rectangle as shown in Fig. 6.3a, or an ellipse.

In contrast, think about constructing U_h using basis functions which have supports on simple triangular domains $\{\Omega_i\}_i$ which are formed by splitting a polygon domain Ω such as the one in Fig. 6.3b. This may enable an approximate solution of the boundary value problem defined over an arbitrary polygon shape to be obtained just by changing the way the basis functions are chosen without the need to change the framework of the Galerkin method. The Galerkin method obtained with these principles is the finite element method. The simple domain chosen in this case is called a finite element.

This section examines in detail the process for solving a one-dimensional Poisson problem using the finite element method. First, define the approximate functions used in the finite element method within the framework of the Galerkin method. Then suppose these approximate functions are defined for the split domains of each finite element. From this, the integration of the weak form can be replaced by the sum of integrations for each finite element domain.

6.2.1 Approximate Functions in Galerkin Method

Consider the finite element method with respect to a one-dimensional Poisson problem (Problem 6.1.1 and its weak form Problem 6.1.2). In the finite element method, the domain $\Omega = (0, 1)$ is split into $(x_0, x_1), (x_1, x_2), \dots, (x_{m-1}, x_m)$ as in Fig. 6.4. Here, x_0, x_1, \dots, x_m are called nodes and $(x_0, x_1), (x_1, x_2), \dots, (x_{m-1}, x_m)$ are called one-dimensional finite elements or the domains of finite elements. The

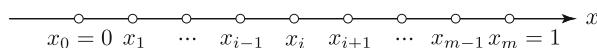


Fig. 6.4 Finite elements and nodes within a one-dimensional domain $\Omega = (0, 1)$

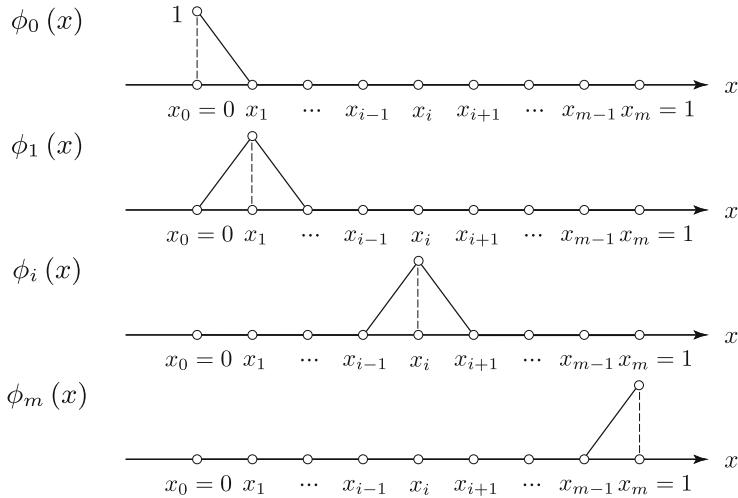


Fig. 6.5 Basis functions ϕ_0, \dots, ϕ_m used in the one-dimensional finite element method

finite elements are numbered and their set represented as $\mathcal{E} = \{1, \dots, m\}$. Moreover, nodes are also numbered and the set of numbers expressed as $\mathcal{N} = \{0, \dots, m\}$.

Select the basis functions in the one-dimensional finite element method with respect to Problem 6.1.1 as a pyramid-shaped function with unit height such as in Fig. 6.5 and defined as

$$\phi_0(x) = \begin{cases} \frac{x_1 - x}{x_1 - x_0} & \text{in } (0, x_1) \\ 0 & \text{in } (x_1, 1) \end{cases},$$

$$\phi_i(x) = \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}} & \text{in } (x_{i-1}, x_i) \\ \frac{x_{i+1} - x}{x_{i+1} - x_i} & \text{in } (x_i, x_{i+1}) \\ 0 & \text{in } (0, x_{i-1}) \cup (x_{i+1}, 1) \end{cases}, \quad \text{for } i \in \mathcal{N},$$

$$\phi_m(x) = \begin{cases} \frac{x - x_{m-1}}{x_m - x_{m-1}} & \text{in } (x_{m-1}, 1) \\ 0 & \text{in } (0, x_{m-1}) \end{cases}.$$

Approximate functions are constructed as

$$u_h(\bar{\mathbf{u}}) = u_0\phi_0 + \sum_{i \in \{1, \dots, m\}} u_i\phi_i = \begin{pmatrix} \bar{u}_D \\ \bar{\mathbf{u}}_N \end{pmatrix} \cdot \begin{pmatrix} \phi_D \\ \phi_N \end{pmatrix} = \bar{\mathbf{u}} \cdot \boldsymbol{\phi}, \quad (6.2.1)$$

$$v_h(\bar{\mathbf{v}}) = v_0\phi_0 + \sum_{i \in \{1, \dots, m\}} v_i\phi_i = \begin{pmatrix} \bar{v}_D \\ \bar{\mathbf{v}}_N \end{pmatrix} \cdot \begin{pmatrix} \phi_D \\ \phi_N \end{pmatrix} = \bar{\mathbf{v}} \cdot \boldsymbol{\phi}, \quad (6.2.2)$$

where $u_0 = \bar{u}_D = u_D$ and $v_0 = \bar{v}_D = 0$. Here, $\bar{u}_N = (u_1, \dots, u_m)^\top$ and $\bar{v}_N = (v_1, \dots, v_m)^\top$ are undetermined multipliers. In this expression, (\cdot) was included to indicate a vector, but \bar{u}_D and \bar{v}_D are one-dimensional vectors for Problem 6.1.1 in which the fundamental boundary condition was given at one node.

Take a look at the characteristics of the approximate functions defined above. First, the fact that the basis functions are continuous functions means that those are included in $H^1(\Omega; \mathbb{R})$ (Sobolev embedding theorem (Theorem 4.3.14)) and satisfy the requirements for substituting into the weak form. The fact that those are first-order polynomials in the finite elements represents that the evaluation of derivatives of ϕ_i and ϕ_j which appear in $a(\phi_i, \phi_j)$ becomes easier. Moreover, the basis functions defined for each node will have the supports only on the finite elements adjacent to the node, hence the domain of integration of $a(\phi_i, \phi_j)$ is limited to each of their finite elements. Furthermore, the unknown multiplier u_i with respect to basis function ϕ_i which is 1 at the node $i \in \mathcal{N}$ matches the node value of approximate function as in Fig. 6.6. From this, \bar{u} and \bar{v} are called nodal value vectors. In this book, the elements \bar{u} and \bar{v} are split into two types, $\bar{u}_D = u_0$ and $\bar{v}_D = v_0$ providing the fundamental boundary condition, called Dirichlet-type nodal value vectors (real numbers in this case) and $\bar{u}_N = (u_1, \dots, u_m)^\top$ and $\bar{v}_N = (v_1, \dots, v_m)^\top$, called Neumann-type nodal value vectors.

6.2.2 Approximate Functions in Finite Element Method

So far, based on the awareness that the finite element method is one form of the Galerkin method, the domain of an approximate function has been taken to be Ω . However, if we focus on the set of basis functions with support on the domain $\Omega_i = (x_{i-1}, x_i)$ for the finite element $i \in \mathcal{E}$, two basis functions $\phi_{i-1}(x)$ and $\phi_i(x)$ with

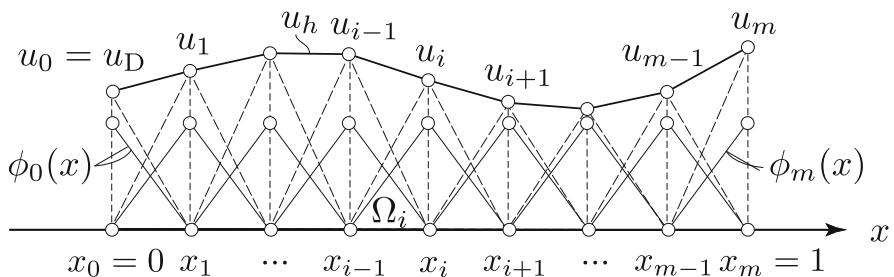
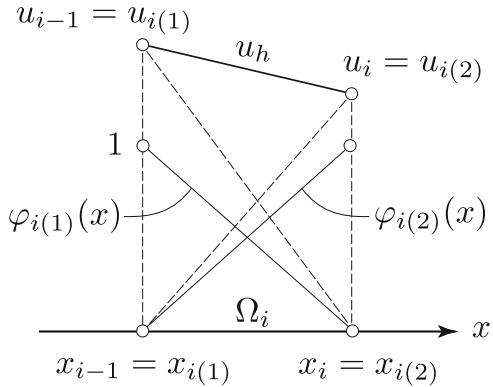


Fig. 6.6 Node value vector \bar{u} in the 1D finite element method

Fig. 6.7 Basis functions $\varphi_{i(1)}, \varphi_{i(2)}$ on a finite element in the 1D finite element method



respect to nodes $i - 1 \in \mathcal{N}$ and $i \in \mathcal{N}$ such as those shown in Fig. 6.7 can be used by rewriting as

$$\varphi_{i(1)}(x) = \phi_{i-1}(x) = \frac{x_{i(2)} - x}{x_{i(2)} - x_{i(1)}}, \quad (6.2.3)$$

$$\varphi_{i(2)}(x) = \phi_i(x) = \frac{x - x_{i(1)}}{x_{i(2)} - x_{i(1)}}, \quad (6.2.4)$$

to define the approximate function as

$$u_h(\bar{u}_i) = (\varphi_{i(1)} \ \varphi_{i(2)}) \begin{pmatrix} u_{i(1)} \\ u_{i(2)} \end{pmatrix} = \boldsymbol{\varphi}_i \cdot \bar{u}_i, \quad (6.2.5)$$

$$v_h(\bar{v}_i) = (\varphi_{i(1)} \ \varphi_{i(2)}) \begin{pmatrix} v_{i(1)} \\ v_{i(2)} \end{pmatrix} = \boldsymbol{\varphi}_i \cdot \bar{v}_i \quad (6.2.6)$$

on Ω_i with respect to all $i \in \mathcal{E}$. In this case, $\boldsymbol{\varphi}_i(x) = (\varphi_{i-1}(x), \varphi_i(x))^\top = (\varphi_{i(1)}(x), \varphi_{i(2)}(x))^\top$ is referred to as basis functions in a finite element. Moreover, in technical books on the finite element method, $\boldsymbol{\varphi}_i(x)$ is generally referred to as a shape function or an interpolation function. However, given that the main topic of this book is the shape optimization problem, the term shape function may bring some confusion. Hence, in this book $\boldsymbol{\varphi}_i(x)$ will be referred to as a basis function in the finite element or basis function when there is no danger of confusion. Moreover, in Eqs. (6.2.5) and (6.2.6),

$$\bar{u}_i = \begin{pmatrix} u_{i-1} \\ u_i \end{pmatrix} = \begin{pmatrix} u_{i(1)} \\ u_{i(2)} \end{pmatrix}, \quad \bar{v}_i = \begin{pmatrix} v_{i-1} \\ v_i \end{pmatrix} = \begin{pmatrix} v_{i(1)} \\ v_{i(2)} \end{pmatrix}$$

is called the element nodal value vector of u and v with respect to finite element $i \in \mathcal{E}$. Hereafter, by using notation $(\cdot)_{i(\alpha)}$ as a function or a value corresponding to

the local node number $\alpha \in \{1, 2\}$ at the finite element $i \in \mathcal{E}$, $u_{i(\alpha)}$ and $v_{i(\alpha)}$ will be referred to as a local node number expression of the finite element $i \in \mathcal{E}$. In addition, since $u_h(\bar{\mathbf{u}}_i)$ and $v_h(\bar{\mathbf{v}}_i)$ of Eqs. (6.2.5) and (6.2.6) are functions $\Omega_i \rightarrow \mathbb{R}$, then note that they can be written as $u_h(\bar{\mathbf{u}}_i)(x)$ and $v_h(\bar{\mathbf{v}}_i)(x)$ with respect to $x \in \Omega_i$. However, the reason that the notations $u_h(\bar{\mathbf{u}}_i)$ and $v_h(\bar{\mathbf{v}}_i)$ are used here is because undetermined multipliers $\bar{\mathbf{u}}_i$ and $\bar{\mathbf{v}}_i$ are variables in the finite element formulation. On the other hand, the basis functions $\varphi_i(x) = (\varphi_{i(1)}(x), \varphi_{i(2)}(x))^\top$ in the finite element are functions of $x \in \Omega_i$ and are constructed using

$$\bar{\mathbf{x}}_i = \begin{pmatrix} x_{i-1} \\ x_i \\ x_{i(2)} \end{pmatrix} = \begin{pmatrix} x_{i-1} \\ x_{i(1)} \\ x_{i(2)} \end{pmatrix}.$$

$\bar{\mathbf{x}}_i$ in this case is called the element node vector with respect to the finite element $i \in \mathcal{E}$.

The basis functions on the finite element defined in this way satisfy

$$\varphi_{i(\alpha)}(x_{i(\beta)}) = \delta_{\alpha\beta} \quad (6.2.7)$$

with respect to $\alpha, \beta \in \{1, 2\}$. Moreover, the equation

$$\sum_{\alpha \in \{1, 2\}} \varphi_{i(\alpha)}(x) = 1 \quad (6.2.8)$$

holds for all $x \in \Omega_i$ with respect to all $i \in \mathcal{E}$. Equation (6.2.7) is a condition for the undetermined multipliers $\bar{\mathbf{u}}_i$ and $\bar{\mathbf{v}}_i$ to represent the node values of the approximate functions. Moreover, Eq. (6.2.8) is a condition for expressing $u = 1$ exactly over Ω_i using $\bar{\mathbf{u}}_i = (1, 1)^\top$.

In order to relate functions $u_h(\bar{\mathbf{u}}_i)$ and $v_h(\bar{\mathbf{v}}_i)$ on $\Omega_i = (x_{i-1}, x_i)$ defined on Eqs. (6.2.5) and (6.2.6) with the total node value vectors $\bar{\mathbf{u}} = (u_0, \dots, u_m)^\top$ and $\bar{\mathbf{v}} = (v_0, \dots, v_m)^\top$, a matrix $\mathbf{Z}_i \in \mathbb{R}^{3 \times (m+1)}$ used as

$$u_h(\bar{\mathbf{u}}_i) = (\varphi_{i(1)} \ \varphi_{i(2)}) \begin{pmatrix} 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 0 \end{pmatrix} \begin{pmatrix} u_0 \\ \vdots \\ u_{i-1} \\ u_i \\ \vdots \\ u_m \end{pmatrix} = \varphi_i \cdot (\mathbf{Z}_i \bar{\mathbf{u}}), \quad (6.2.9)$$

$$v_h(\bar{\mathbf{v}}_i) = \varphi_i \cdot (\mathbf{Z}_i \bar{\mathbf{v}}) \quad (6.2.10)$$

is introduced. Such a \mathbf{Z}_i is called Boolean matrix.

6.2.3 Discretized Equations

The approximate functions for each finite element were constructed as Eqs. (6.2.9) and (6.2.10). Therefore, it is possible to substitute these into the weak form (Problem 6.1.2) of the one-dimensional Poisson problem and to obtain the discretized equation with $\bar{\mathbf{u}}_N$ as an unknown.

Substituting $u_h(\bar{\mathbf{u}})$ and $v_h(\bar{\mathbf{v}})$ of Eqs. (6.2.1) and (6.2.2) into the weak form, we get

$$a(u_h(\bar{\mathbf{u}}), v_h(\bar{\mathbf{v}})) = l(v_h(\bar{\mathbf{v}})). \quad (6.2.11)$$

Here, the left-hand side of Eq. (6.2.11) can split the domain of integration for each element and can be written as

$$\begin{aligned} a(u_h(\bar{\mathbf{u}}), v_h(\bar{\mathbf{v}})) &= \sum_{i \in \{1, \dots, m\}} \int_{x_{i(1)}}^{x_{i(2)}} \frac{d u_h}{d x}(\bar{\mathbf{u}}_i) \frac{d v_h}{d x}(\bar{\mathbf{v}}_i) dx \\ &= \sum_{i \in \{1, \dots, m\}} a_i(u_h(\bar{\mathbf{u}}_i), v_h(\bar{\mathbf{v}}_i)). \end{aligned} \quad (6.2.12)$$

Each term on the right-hand side of Eq. (6.2.12) can be summarized using $u_h(\bar{\mathbf{u}}_i)$ and $v_h(\bar{\mathbf{v}}_i)$ of Eqs. (6.2.9) and (6.2.10) as

$$\begin{aligned} a_i(u_h(\bar{\mathbf{u}}_i), v_h(\bar{\mathbf{v}}_i)) &= (v_{i(1)} \ v_{i(2)}) \\ &= \begin{pmatrix} v_{i(1)} & v_{i(2)} \end{pmatrix} \\ &\quad \times \begin{pmatrix} \int_{x_{i(1)}}^{x_{i(2)}} \frac{d \varphi_{i(1)}}{d x} \frac{d \varphi_{i(1)}}{d x} dx & \int_{x_{i(1)}}^{x_{i(2)}} \frac{d \varphi_{i(1)}}{d x} \frac{d \varphi_{i(2)}}{d x} dx \\ \int_{x_{i(1)}}^{x_{i(2)}} \frac{d \varphi_{i(2)}}{d x} \frac{d \varphi_{i(1)}}{d x} dx & \int_{x_{i(1)}}^{x_{i(2)}} \frac{d \varphi_{i(2)}}{d x} \frac{d \varphi_{i(2)}}{d x} dx \end{pmatrix} \begin{pmatrix} u_{i(1)} \\ u_{i(2)} \end{pmatrix} \\ &= (v_{i(1)} \ v_{i(2)}) \begin{pmatrix} a_i(\varphi_{i(1)}, \varphi_{i(1)}) & a_i(\varphi_{i(1)}, \varphi_{i(2)}) \\ a_i(\varphi_{i(2)}, \varphi_{i(1)}) & a_i(\varphi_{i(2)}, \varphi_{i(2)}) \end{pmatrix} \begin{pmatrix} u_{i(1)} \\ u_{i(2)} \end{pmatrix} \\ &= \bar{\mathbf{v}}_i \cdot (\bar{\mathbf{A}}_i \bar{\mathbf{u}}_i) = \bar{\mathbf{v}} \cdot (\mathbf{Z}_i^\top \bar{\mathbf{A}}_i \mathbf{Z}_i \bar{\mathbf{u}}) = \bar{\mathbf{v}} \cdot (\tilde{\mathbf{A}}_i \bar{\mathbf{u}}). \end{aligned} \quad (6.2.13)$$

Here, $\bar{\mathbf{A}}_i = (\bar{a}_{i(\alpha\beta)})_{\alpha\beta} \in \mathbb{R}^{2 \times 2}$ is called the coefficient matrix of the finite element $i \in \mathcal{E}$. Let $\tilde{\mathbf{A}}_i \in \mathbb{R}^{(m+1) \times (m+1)}$ be a matrix which has been expanded with zero added to go with the total nodal value vectors. It should be noted that in contrast to the element nodal value vectors $\bar{\mathbf{u}}_i$ and $\bar{\mathbf{v}}_i$ of the finite element $i \in \mathcal{E}$ being elements of \mathbb{R}^2 , the total nodal value vectors $\bar{\mathbf{u}}$ and $\bar{\mathbf{v}}$ are elements of \mathbb{R}^{m+1} .

If Eqs. (6.2.3) and (6.2.4) are used to calculate $\bar{\mathbf{A}}_i$, one obtains

$$\begin{aligned}\bar{a}_{i(11)} &= \int_{x_{i(1)}}^{x_{i(2)}} \frac{d\varphi_{i(1)}}{dx} \frac{d\varphi_{i(1)}}{dx} dx = \frac{1}{(x_{i(2)} - x_{i(1)})^2} \int_{x_{i(1)}}^{x_{i(2)}} (-1)^2 dx \\ &= \frac{1}{x_{i(2)} - x_{i(1)}}, \\ \bar{a}_{i(12)} &= \int_{x_{i(1)}}^{x_{i(2)}} \frac{d\varphi_{i(1)}}{dx} \frac{d\varphi_{i(2)}}{dx} dx = \frac{1}{(x_{i(2)} - x_{i(1)})^2} \int_{x_{i(1)}}^{x_{i(2)}} 1 \cdot (-1) dx \\ &= \frac{-1}{x_{i(2)} - x_{i(1)}}, \\ \bar{a}_{i(21)} &= \bar{a}_{i(12)}, \\ \bar{a}_{i(22)} &= \int_{x_{i(1)}}^{x_{i(2)}} \frac{d\varphi_{i(2)}}{dx} \frac{d\varphi_{i(2)}}{dx} dx = \frac{1}{(x_{i(2)} - x_{i(1)})^2} \int_{x_{i(1)}}^{x_{i(2)}} 1^2 dx \\ &= \frac{1}{x_{i(2)} - x_{i(1)}}\end{aligned}$$

and

$$\bar{\mathbf{A}}_i = \begin{pmatrix} \bar{a}_{i(11)} & \bar{a}_{i(12)} \\ \bar{a}_{i(21)} & \bar{a}_{i(22)} \end{pmatrix} = \frac{1}{x_{i(2)} - x_{i(1)}} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \quad (6.2.14)$$

On the other hand, the right-hand side of Eq. (6.2.11) can also be split into each element as

$$\begin{aligned}l(v_h(\bar{\mathbf{v}})) &= \sum_{i \in \{1, \dots, m\}} \int_{x_{i(1)}}^{x_{i(2)}} b v_h(\bar{\mathbf{v}}_i) dx + p_N v_h(\bar{\mathbf{v}}_m) \\ &= \sum_{i \in \{1, \dots, m\}} l_i(v_h(\bar{\mathbf{v}}_i)).\end{aligned} \quad (6.2.15)$$

Here, for the finite element $i \in \{1, \dots, m-1\}$ and m , let

$$\begin{aligned}l_i(v_h(\bar{\mathbf{v}}_i)) &= (v_{i(1)} \ v_{i(2)}) \begin{pmatrix} \int_{x_{i(1)}}^{x_{i(2)}} b \varphi_{i(1)} dx \\ \int_{x_{i(1)}}^{x_{i(2)}} b \varphi_{i(2)} dx \end{pmatrix} = (v_{i(1)} \ v_{i(2)}) \begin{pmatrix} \bar{b}_{i(1)} \\ \bar{b}_{i(2)} \end{pmatrix} \\ &= \bar{\mathbf{v}}_i \cdot \bar{\mathbf{b}}_i = \bar{\mathbf{v}} \cdot (\mathbf{Z}_i^\top \bar{\mathbf{b}}_i) = \bar{\mathbf{v}} \cdot \tilde{\mathbf{b}}_i \\ &= \bar{\mathbf{v}}_i \cdot \bar{\mathbf{l}}_i = \bar{\mathbf{v}} \cdot (\mathbf{Z}_i^\top \bar{\mathbf{l}}_i) = \bar{\mathbf{v}} \cdot \tilde{\mathbf{l}}_i,\end{aligned} \quad (6.2.16)$$

$$\begin{aligned}
l_m(v_h(\bar{\mathbf{v}}_m)) &= (v_{m(1)} \ v_{m(2)}) \left(\left(\begin{array}{l} \int_{x_{m(1)}}^{x_{m(2)}} b \varphi_{i(1)} dx \\ \int_{x_{m(1)}}^{x_{m(2)}} b \varphi_{i(2)} dx \end{array} \right) + \begin{pmatrix} 0 \\ p_N \end{pmatrix} \right) \\
&= (v_{m(1)} \ v_{m(2)}) \left(\begin{pmatrix} \bar{b}_{m(1)} \\ \bar{b}_{m(2)} \end{pmatrix} + \begin{pmatrix} \bar{p}_{m(1)} \\ \bar{p}_{m(2)} \end{pmatrix} \right) \\
&= \bar{\mathbf{v}}_m \cdot (\bar{\mathbf{b}}_m + \bar{\mathbf{p}}_m) = \bar{\mathbf{v}} \cdot \left\{ \mathbf{Z}_m^\top (\bar{\mathbf{b}}_m + \bar{\mathbf{p}}_m) \right\} = \bar{\mathbf{v}} \cdot (\tilde{\mathbf{b}}_m + \tilde{\mathbf{p}}_m) \\
&= \bar{\mathbf{v}}_m \cdot \bar{\mathbf{l}}_m = \bar{\mathbf{v}} \cdot (\mathbf{Z}_m^\top \bar{\mathbf{l}}_m) = \bar{\mathbf{v}} \cdot \bar{\mathbf{l}}_m,
\end{aligned} \tag{6.2.17}$$

respectively. Here, with respect to $i \in \mathcal{E} = \{1, \dots, m\}$, $\bar{\mathbf{l}}_i$ is called the known term vector of the finite element i . $\bar{\mathbf{b}}_i$ and $\bar{\mathbf{p}}_i$ represent the components of the known term vector constructed from b and p_N respectively. $\bar{\mathbf{l}}_i$, $\tilde{\mathbf{b}}_i$ and $\tilde{\mathbf{p}}_i$ are $\bar{\mathbf{l}}_i$, $\bar{\mathbf{b}}_i$ and $\bar{\mathbf{p}}_i$ respectively expanded by adding in 0 to match the total nodal value vectors.

When b is a constant function, $\bar{\mathbf{b}}_i = (\bar{b}_{i(1)}, \bar{b}_{i(2)})^\top$ is calculated as

$$\begin{aligned}
\bar{b}_{i(1)} &= b \int_{x_{i(1)}}^{x_{i(2)}} \varphi_{i(1)} dx = b \int_{x_{i(1)}}^{x_{i(2)}} \frac{x_{i(2)} - x}{x_{i(2)} - x_{i(1)}} dx = b \frac{x_{i(2)} - x_{i(1)}}{2}, \\
\bar{b}_{i(2)} &= b \int_{x_{i(1)}}^{x_{i(2)}} \varphi_{i(2)} dx = b \int_{x_{i(1)}}^{x_{i(2)}} \frac{x - x_{i(1)}}{x_{i(2)} - x_{i(1)}} dx = b \frac{x_{i(2)} - x_{i(1)}}{2}
\end{aligned}$$

and is obtained as

$$\bar{\mathbf{b}}_i = b \frac{x_{i(2)} - x_{i(1)}}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{6.2.18}$$

Here, if Eq. (6.2.12) with Eq. (6.2.13) substituted in, and Eq. (6.2.15) with Eqs. (6.2.16) and (6.2.17) substituted in are substituted into the weak form (Eq. (6.2.11)), we get

$$\bar{\mathbf{v}} \cdot \sum_{i \in \{1, \dots, m\}} (\tilde{\mathbf{A}}_i \bar{\mathbf{u}}) = \bar{\mathbf{v}} \cdot \sum_{i \in \{1, \dots, m\}} \bar{\mathbf{l}}_i,$$

which can be rewritten as

$$\bar{\mathbf{v}} \cdot (\bar{\mathbf{A}} \bar{\mathbf{u}}) = \bar{\mathbf{v}} \cdot \bar{\mathbf{l}}. \tag{6.2.19}$$

In other words we set

$$\bar{\mathbf{A}} = \sum_{i \in \{1, \dots, m\}} \tilde{\mathbf{A}}_i \in \mathbb{R}^{(m+1) \times (m+1)},$$

$$\bar{\mathbf{l}} = \sum_{i \in \{1, \dots, m\}} \tilde{\mathbf{l}}_i = \sum_{i \in \{1, \dots, m\}} \tilde{\mathbf{b}}_i + \tilde{\mathbf{p}}_m = \bar{\mathbf{b}} + \bar{\mathbf{p}} \in \mathbb{R}^{m+1}.$$

$\bar{\mathbf{A}}$ and $\bar{\mathbf{l}}$ are called the total coefficient matrix and total known term vector, respectively. Moreover, $\bar{\mathbf{b}}$ and $\bar{\mathbf{p}}$ are called total nodal value vectors of b and p_N , respectively.

Equation (6.2.19) gives the weak form but there were no fundamental boundary conditions assumed with respect to u_h and v_h . Hence, substitute in the fundamental boundary conditions into Eq. (6.2.19). $u_0 = \bar{u}_D = u_D$ (\bar{u}_D is the node value of the fundamental boundary condition, u_D is the known value of boundary value problem) and $v_0 = \bar{v}_D = 0$. Substituting these into Eq. (6.2.19) gives

$$(0 \mid v_1 \dots v_m) \left(\begin{array}{c|ccccc} \bar{a}_{00} & \bar{a}_{01} & \dots & \bar{a}_{0m} \\ \bar{a}_{10} & \bar{a}_{11} & \dots & \bar{a}_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{a}_{m0} & \bar{a}_{m1} & \dots & \bar{a}_{mm} \end{array} \right) \begin{pmatrix} u_D \\ u_1 \\ \vdots \\ u_m \end{pmatrix} - \begin{pmatrix} l_0 \\ l_1 \\ \vdots \\ l_m \end{pmatrix} \right) \\ = (0 \mid \bar{v}_N^\top) \left(\begin{pmatrix} \bar{A}_{DD} & \bar{A}_{DN} \\ \bar{A}_{ND} & \bar{A}_{NN} \end{pmatrix} \begin{pmatrix} \bar{u}_D \\ \bar{u}_N \end{pmatrix} - \begin{pmatrix} \bar{l}_D \\ \bar{l}_N \end{pmatrix} \right) = 0. \quad (6.2.20)$$

Rearranging Eq. (6.2.20), we get

$$\begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix} \cdot \left(\begin{pmatrix} \bar{a}_{11} & \dots & \bar{a}_{1m} \\ \vdots & \ddots & \vdots \\ \bar{a}_{m1} & \dots & \bar{a}_{mm} \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix} - \begin{pmatrix} l_1 \\ \vdots \\ l_m \end{pmatrix} + \begin{pmatrix} u_D \bar{a}_{10} \\ \vdots \\ u_D \bar{a}_{m0} \end{pmatrix} \right) \\ = \bar{v}_N^\top (\bar{A}_{NN} \bar{u}_N - \bar{l}_N + \bar{u}_D \bar{A}_{ND}) = 0.$$

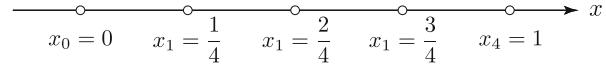
\bar{v}_N is arbitrary, so it can be written as

$$\bar{A}_{NN} \bar{u}_N = \bar{l}_N - \bar{u}_D \bar{A}_{ND} = \hat{l}. \quad (6.2.21)$$

Equation (6.2.21) is a simultaneous linear equation with respect to unknown vector \bar{u}_N and is called the discretized equation of the finite element method. Solving Eq. (6.2.21) for \bar{u}_N gives

$$\bar{u}_N = \bar{A}_{NN}^{-1} \hat{l}. \quad (6.2.22)$$

Fig. 6.8 Finite element mesh when $m = 4$



From this, the finite element solution $u_h(\bar{u})$ can be obtained by substituting $\bar{u} = (\bar{u}_D, \bar{u}_N^\top)^\top$ into Eq. (6.2.1). Moreover, the finite element solution $u_h(\bar{u}_i)$ on Ω_i on the finite element $i \in \mathcal{E}$ domain can be found by Eq. (6.2.9).

6.2.4 Exercise Problem

We solve a one-dimensional Poisson problem using the finite element method in the following exercise.

Exercise 6.2.1 (Finite Element Method for 1D Poisson Problem) Let b be a constant function as in Problem 6.1.1. Here, show the system of linear equations for seeking the approximate solution in the aforementioned one-dimensional finite element method using the finite element mesh shown in Fig. 6.8. Moreover, obtain the approximate solution when we set $b = 1$, $u_D = 0$ and $p_N = 0$. \square

Answer Let the size of the finite element be $h = 1/4$. From Eqs. (6.2.14), (6.2.18) and (6.2.17), we get

$$\bar{A}_i = \frac{1}{h} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad \bar{b}_i = \frac{hb}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \bar{p}_4 = \begin{pmatrix} 0 \\ p_N \end{pmatrix}.$$

Expanding \bar{A}_1 and \bar{b}_1 to match the node value vectors gives

$$\tilde{A}_1 = \frac{1}{h} \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \tilde{b}_1 = \frac{hb}{2} \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

If \tilde{A}_2 and \tilde{b}_2 are overlapped with \tilde{A}_1 and \tilde{b}_1 respectively, we obtain

$$\tilde{A}_1 + \tilde{A}_2 = \frac{1}{h} \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 1 & 1 & -1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \tilde{b}_1 + \tilde{b}_2 = \frac{hb}{2} \begin{pmatrix} 1 \\ 1+1 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

Similarly, overlapping $\tilde{\mathbf{A}}_3$ and $\tilde{\mathbf{b}}_3$, $\tilde{\mathbf{A}}_4$ and $\tilde{\mathbf{b}}_4$ as well as $\tilde{\mathbf{p}}_4$ gives

$$\bar{\mathbf{A}} = \sum_{i \in \{1, \dots, 4\}} \tilde{\mathbf{A}}_i = \frac{1}{h} \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix},$$

$$\bar{\mathbf{l}} = \sum_{i \in \{1, \dots, 4\}} \tilde{\mathbf{b}}_i + \tilde{\mathbf{p}}_4 = \frac{hb}{2} \begin{pmatrix} 1 \\ 2 \\ 2 \\ 2 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ p_N \end{pmatrix}.$$

Substituting fundamental boundary conditions $u_0 = \bar{u}_D = u_D$ and $v_0 = \bar{v}_D = 0$ into Eq. (6.2.19) gives

$$\begin{pmatrix} 0 \\ v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} \cdot \left(\frac{1}{h} \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} u_D \\ u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} - \frac{hb}{2} \begin{pmatrix} 1 \\ 2 \\ 2 \\ 2 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ p_N \end{pmatrix} \right) = 0.$$

Rearranging this equation gives

$$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} \cdot \left(\frac{1}{h} \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} - \frac{hb}{2} \begin{pmatrix} 2 \\ 2 \\ 2 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} - \frac{1}{h} \begin{pmatrix} u_D \\ 0 \\ 0 \\ 0 \end{pmatrix} \right) = 0.$$

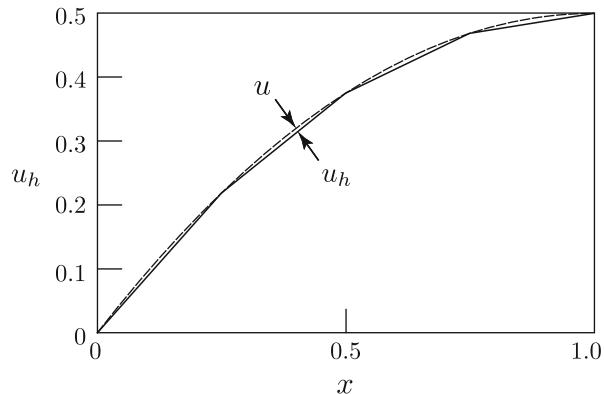
(v_1, v_2, v_3, v_4) is arbitrary, so

$$\frac{1}{h} \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \frac{hb}{2} \begin{pmatrix} 2 \\ 2 \\ 2 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \frac{1}{h} \begin{pmatrix} u_D \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

can be obtained. This equation is a simultaneous linear equation such as

$$\bar{\mathbf{A}}_{NN} \bar{\mathbf{u}}_N = \hat{\mathbf{l}}$$

Fig. 6.9 Exact solution u and approximate solution u_h of Exercise 6.2.1



with respect to \bar{u}_N . Here, when $b = 1$, $u_D = 0$, $p_N = 0$, we get

$$\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \frac{1}{4^2} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1/2 \end{pmatrix} = \begin{pmatrix} 7/32 \\ 3/8 \\ 15/32 \\ 1/2 \end{pmatrix}.$$

On the other hand, the exact solution is

$$u = -\frac{1}{2}x^2 + x.$$

Figure 6.9 shows a comparison between the numerical solution u_h and the exact solution u . \square

Change the boundary conditions of Exercise 6.2.1 and consider the following problem.

Exercise 6.2.2 (Dirichlet Problem of 1D Poisson Problem) Consider a problem seeking $u : (0, 1) \rightarrow \mathbb{R}$ satisfying

$$-\frac{d^2u}{dx^2} = b \quad \text{in } (0, 1), \quad u(0) = u_{D0}, \quad u(1) = u_{D1}$$

when $b, u_{D0}, u_{D1}, p_N \in \mathbb{R}$ are given. Use the mesh for the finite elements of Fig. 6.8 in order to show the simultaneous linear equations when seeking the approximate solution using the finite element method. Moreover, obtain the numerical solution when $b = 1$ and $u_{D0} = u_{D1} = 0$. \square

Answer The calculation of \bar{A} is similar to Exercise 6.2.1. Moreover $\bar{l} = \sum_{i \in \{1, \dots, 4\}} \bar{b}_i$. If the fundamental boundary conditions $u_0 = \bar{u}_{D0} = u_{D0}$, $u_4 = \bar{u}_{D4} = u_{D1}$, $v_0 = \bar{v}_{D0} = 0$ and $v_4 = \bar{v}_{D4} = 0$ are substituted into

Eq. (6.2.19), we get

$$\begin{pmatrix} 0 \\ v_1 \\ v_2 \\ v_3 \\ 0 \end{pmatrix} \cdot \left(\frac{1}{h} \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} u_{D0} \\ u_1 \\ u_2 \\ u_3 \\ u_{D1} \end{pmatrix} - \frac{hb}{2} \begin{pmatrix} 1 \\ 2 \\ 2 \\ 2 \\ 1 \end{pmatrix} \right) = 0.$$

Rearranging this equation gives

$$(v_1 \ v_2 \ v_3) \left(\frac{1}{h} \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} - \frac{h}{2} \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} - \frac{1}{h} \begin{pmatrix} u_{D0} \\ 0 \\ u_{D1} \end{pmatrix} \right) = 0.$$

Since (v_1, v_2, v_3) is arbitrary,

$$\frac{1}{h} \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \frac{h}{2} \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} + \frac{1}{h} \begin{pmatrix} u_{D0} \\ 0 \\ u_{D1} \end{pmatrix}$$

is obtained. These equations are simultaneous linear equations like

$$\bar{A}_{NN} \bar{u}_N = \hat{l}.$$

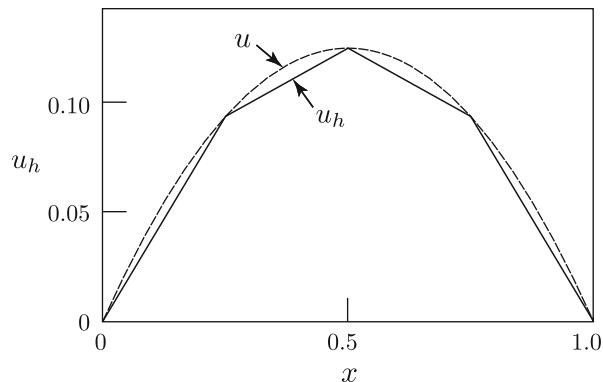
When $b = 1$ and $u_{D0} = u_{D1} = 0$, we get

$$\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \frac{1}{4^3} \begin{pmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3/32 \\ 1/8 \\ 3/32 \end{pmatrix}.$$

Figure 6.10 shows the comparison between the numerical solution u_h and the exact solution $u = \frac{1}{2}x(x-1)$. \square

The following can be said to be one of the characteristics of the finite element method from Exercise 6.2.1 and Exercise 6.2.2. The Galerkin method seen in Sect. 6.1 required a change of basis functions for a change in fundamental boundary conditions. However, in the finite element method, fundamental boundary conditions can be changed easily as boundary conditions are substituted in after seeking the coefficient matrix and known term vector.

Fig. 6.10 Exact solution u and approximate solution u_h of Exercise 6.2.2



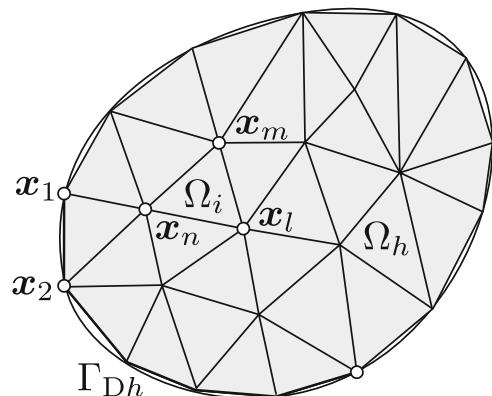
6.3 Two-Dimensional Finite Element Method

Next, we consider a finite element method with respect to a two-dimensional Poisson problem (Problem 6.1.6 with $d = 2$). Here, approximate functions used in the finite element method within the framework of the Galerkin method are also defined. After that, their approximate functions will be viewed as approximate functions defined for each split finite element domain.

6.3.1 Approximate Functions in Galerkin Method

With reference to Fig. 6.11, a two-dimensional domain Ω and Dirichlet boundary Γ_D are assumed to be approximated by a polygonal domain Ω_h and line graph Γ_{Dh} , respectively. Furthermore, Ω_h is split into a set of triangle domains $\{\Omega_i\}_i$. In this case, Ω_i is called the domain of triangular finite elements and the set of finite

Fig. 6.11 Triangular finite elements and nodes in a 2D domain Ω



element numbers i is denoted by \mathcal{E} . Here, we assume that there are no overlapping triangular domains Ω_i for all $i \in \mathcal{E}$ and there are no vertices on the boundary of triangular domains other than the vertices of Ω_i as in Fig. 6.12.

Moreover, the vertices $\mathbf{x}_j = (x_{j1}, x_{j2})^\top$ of triangles are called nodes and the set of node numbers j is denoted as \mathcal{N} . Furthermore, \mathcal{N} is split into two sets that are the set \mathcal{N}_D of node numbers on Γ_{Dh} and the set $\mathcal{N}_N = \mathcal{N} \setminus \mathcal{N}_D$ of the other node numbers. \mathcal{N} is reordered so that \mathcal{N}_D comes first.

With respect to a triangular finite element mesh such as this, basis functions in the two-dimensional finite element method with respect to Problem 6.1.6 are defined by a pyramidal function with a unit height such as shown in Fig. 6.13 with respect to a node $j \in \mathcal{N}$. In other words, it is assumed that ϕ_j is a first-order polynomial with support on the finite elements having a node j on the vertex and is a continuous function taking the value 1 at node j and the value 0 at other nodes. These characteristics, as seen in the one-dimensional finite element method, lead to the conclusion that integration after substitution into the weak form is convenient and the unknown multipliers in the approximate functions become node values as shown next.

Using these basis functions, the finite element method sets the approximate functions to be

$$u_h(\bar{\mathbf{u}}) = \sum_{j \in \mathcal{N}_D} u_j \phi_j + \sum_{j \in \mathcal{N}_N} u_j \phi_j = \begin{pmatrix} \bar{\mathbf{u}}_D \\ \bar{\mathbf{u}}_N \end{pmatrix} \cdot \begin{pmatrix} \phi_D \\ \phi_N \end{pmatrix} = \bar{\mathbf{u}} \cdot \phi, \quad (6.3.1)$$

$$v_h(\bar{\mathbf{v}}) = \sum_{j \in \mathcal{N}_D} v_j \phi_j + \sum_{j \in \mathcal{N}_N} v_j \phi_j = \begin{pmatrix} \bar{\mathbf{v}}_D \\ \bar{\mathbf{v}}_N \end{pmatrix} \cdot \begin{pmatrix} \phi_D \\ \phi_N \end{pmatrix} = \bar{\mathbf{v}} \cdot \phi. \quad (6.3.2)$$

Here, from the fact that $\bar{\mathbf{u}}$ and $\bar{\mathbf{v}}$ represent the node values of the approximate functions u_h and v_h respectively, $\bar{\mathbf{u}}$ and $\bar{\mathbf{v}}$ are called nodal value vectors. Moreover, $\bar{\mathbf{u}}_D = (u_D(\mathbf{x}_j))_{j \in \mathcal{N}_D}$ and $\bar{\mathbf{v}}_D = \mathbf{0}_{\mathbb{R}^{|\mathcal{N}_D|}}$ are called Dirichlet-type nodal value vectors

Fig. 6.12 Counterexample of triangular finite element and nodes

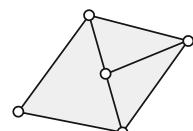
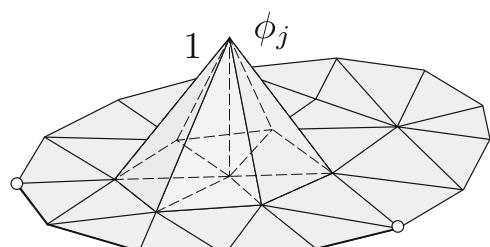


Fig. 6.13 Basis function ϕ_j



and $\bar{\mathbf{u}}_N = (u_j)_{j \in \mathcal{N}_N}$ and $\bar{\mathbf{v}}_N = (v_j)_{j \in \mathcal{N}_N}$ are called Neumann-type nodal value vectors.

6.3.2 Approximate Functions in Finite Element Method

In the finite element method, the basis functions ϕ defined on Ω_h can be rewritten as basis functions defined on Ω_i with respect to all $i \in \mathcal{E}$. Based on that, let the approximate function on Ω_i be

$$u_h(\bar{\mathbf{u}}_i) = (\varphi_{i(1)} \ \varphi_{i(2)} \ \varphi_{i(3)}) \begin{pmatrix} u_{i(1)} \\ u_{i(2)} \\ u_{i(3)} \end{pmatrix} = \boldsymbol{\varphi}_i \cdot \bar{\mathbf{u}}_i, \quad (6.3.3)$$

$$v_h(\bar{\mathbf{v}}_i) = (\varphi_{i(1)} \ \varphi_{i(2)} \ \varphi_{i(3)}) \begin{pmatrix} v_{i(1)} \\ v_{i(2)} \\ v_{i(3)} \end{pmatrix} = \boldsymbol{\varphi}_i \cdot \bar{\mathbf{v}}_i. \quad (6.3.4)$$

Here, when the three node numbers of the finite element $i \in \mathcal{E}$ are l, m and $n \in \mathcal{N}$, $\boldsymbol{\varphi}_i = (\varphi_l, \varphi_m, \varphi_n)^\top = (\varphi_{i(1)}, \varphi_{i(2)}, \varphi_{i(3)})^\top : \Omega_i \rightarrow \mathbb{R}^3$ are referred to as basis functions in the finite element. Moreover,

$$\begin{aligned} \bar{\mathbf{x}}_i &= \begin{pmatrix} \mathbf{x}_l \\ \mathbf{x}_m \\ \mathbf{x}_n \end{pmatrix} = \begin{pmatrix} \mathbf{x}_{i(1)} \\ \mathbf{x}_{i(2)} \\ \mathbf{x}_{i(3)} \end{pmatrix}, \\ \bar{\mathbf{u}}_i &= \begin{pmatrix} u_l \\ u_m \\ u_n \end{pmatrix} = \begin{pmatrix} u_{i(1)} \\ u_{i(2)} \\ u_{i(3)} \end{pmatrix}, \quad \bar{\mathbf{v}}_i = \begin{pmatrix} v_l \\ v_m \\ v_n \end{pmatrix} = \begin{pmatrix} v_{i(1)} \\ v_{i(2)} \\ v_{i(3)} \end{pmatrix} \end{aligned}$$

are called the element node vector and element nodal value vectors with respect to u and v for the finite element $i \in \mathcal{E}$. The suffix $\alpha \in \{1, 2, 3\}$ used in $\mathbf{x}_{i(\alpha)}$, $u_{i(\alpha)}$ and $v_{i(\alpha)}$ is called the local node number of the finite element $i \in \mathcal{E}$. Figure 6.14 shows how the approximate function u_h in the finite element $i \in \mathcal{E}$ is constructed from $\boldsymbol{\varphi}_i$ and $\bar{\mathbf{u}}_i$.

Basis functions $\boldsymbol{\varphi}_i$ on $i \in \mathcal{E}$ constructed in this way satisfy

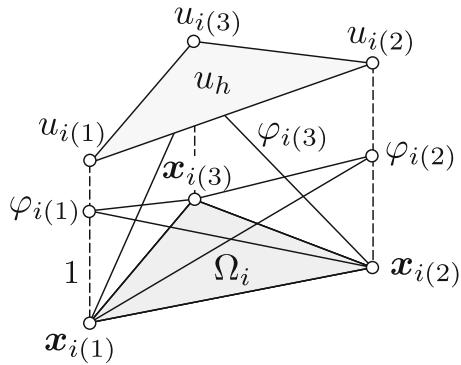
$$\varphi_{i(\alpha)}(\mathbf{x}_{i(\beta)}) = \delta_{\alpha\beta} \quad (6.3.5)$$

with respect to $\alpha, \beta \in \{1, 2, 3\}$. Moreover,

$$\sum_{\alpha \in \{1, 2, 3\}} \varphi_{i(\alpha)}(\mathbf{x}) = 1$$

holds on all points $\mathbf{x} \in \Omega_i$.

Fig. 6.14 Basis functions $\varphi_{i(1)}, \varphi_{i(2)}, \varphi_{i(3)}$ in triangular finite element $i \in \mathcal{E}$



Here, obtain the equations of basis functions $\varphi_{i(1)}, \varphi_{i(2)}$ and $\varphi_{i(3)}$ on the finite element $i \in \mathcal{E}$. From the fact that $\varphi_{i(\alpha)}$ for $\alpha \in \{1, 2, 3\}$ is a linear equation with respect to $\mathbf{x} = (x_1, x_2)^\top \in \Omega_i$, it is a complete first-order polynomial constructed of three unknown multipliers. Let this be

$$\varphi_{i(\alpha)} = \zeta_\alpha + \eta_\alpha x_1 + \theta_\alpha x_2. \quad (6.3.6)$$

The undetermined multipliers $\zeta_\alpha, \eta_\alpha$ and θ_α can be determined by giving the values of $\varphi_{i(\alpha)}$ at three nodes. Their values can be given by Eq. (6.3.5). In other words, with respect to $\beta \in \{1, 2, 3\}$, it can be determined by

$$\varphi_{i(\alpha)}(\mathbf{x}_{i(\beta)}) = \zeta_\alpha + \eta_\alpha x_{i(\beta)1} + \theta_\alpha x_{i(\beta)2} = \delta_{\alpha\beta}.$$

This equation can be expanded as

$$\begin{pmatrix} 1 & x_{i(1)1} & x_{i(1)2} \\ 1 & x_{i(2)1} & x_{i(2)2} \\ 1 & x_{i(3)1} & x_{i(3)2} \end{pmatrix} \begin{pmatrix} \zeta_1 & \zeta_2 & \zeta_3 \\ \eta_1 & \eta_2 & \eta_3 \\ \theta_1 & \theta_2 & \theta_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Here, solving for the undetermined multipliers gives

$$\begin{pmatrix} \zeta_1 & \zeta_2 & \zeta_3 \\ \eta_1 & \eta_2 & \eta_3 \\ \theta_1 & \theta_2 & \theta_3 \end{pmatrix} = \frac{1}{\gamma} \begin{pmatrix} x_{i(2)1}x_{i(3)2} - x_{i(3)1}x_{i(2)2} & x_{i(2)2} - x_{i(3)2} & x_{i(3)1} - x_{i(2)1} \\ x_{i(3)1}x_{i(1)2} - x_{i(1)1}x_{i(3)2} & x_{i(1)1}x_{i(2)2} - x_{i(2)1}x_{i(1)2} & x_{i(1)2} - x_{i(2)2} \\ x_{i(1)2} - x_{i(3)2} & x_{i(1)1} - x_{i(3)1} & x_{i(2)1} - x_{i(1)1} \end{pmatrix}, \quad (6.3.7)$$

where

$$\begin{aligned}\gamma &= \begin{vmatrix} x_{i(1)1} & x_{i(1)2} & 1 \\ x_{i(2)1} & x_{i(2)2} & 1 \\ x_{i(3)1} & x_{i(3)2} & 1 \end{vmatrix} \\ &= x_{i(1)1} (x_{i(2)2} - x_{i(3)2}) + x_{i(2)1} (x_{i(3)2} - x_{i(1)2}) \\ &\quad + x_{i(3)1} (x_{i(1)2} - x_{i(2)2}).\end{aligned}\quad (6.3.8)$$

If the three nodes $\mathbf{x}_{i(1)}$, $\mathbf{x}_{i(2)}$ and $\mathbf{x}_{i(3)}$ of the triangular finite element are chosen so that they are anti-clockwise, then γ is equal to twice the area $|\Omega_i|$ of the triangle Ω_i (Practice 6.3).

The basis functions $\varphi_{i(1)}$, $\varphi_{i(2)}$ and $\varphi_{i(3)}$ of the finite element were obtained by substituting Eqs. (6.3.7) and (6.3.8) into Eq. (6.3.6). Using these, the approximate functions $u_h(\bar{\mathbf{u}}_i)$ and $v_h(\bar{\mathbf{v}}_i)$ defined by Eqs. (6.3.3) and (6.3.4) can be written as

$$u_h(\bar{\mathbf{u}}_i) = (\varphi_{i(1)} \ \varphi_{i(2)} \ \varphi_{i(3)}) \begin{pmatrix} 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_l \\ u_m \\ u_n \\ \vdots \\ u_{|\mathcal{N}|} \end{pmatrix} = \varphi_i \cdot (\mathbf{Z}_i \bar{\mathbf{u}}), \quad (6.3.9)$$

$$v_h(\bar{\mathbf{v}}_i) = \varphi_i \cdot (\mathbf{Z}_i \bar{\mathbf{v}}). \quad (6.3.10)$$

Here, \mathbf{Z}_i is a Boolean matrix which links the total nodal value vector $\bar{\mathbf{u}}$ and the nodal value vector $\bar{\mathbf{u}}_i = (u_l, u_m, u_n)^\top$ of the finite element $i \in \mathcal{E}$.

6.3.3 Discretized Equations

Approximate functions $u_h(\bar{\mathbf{u}}_i)$ and $v_h(\bar{\mathbf{v}}_i)$ on Ω_i have been defined, so we shall substitute Eqs. (6.3.9) and (6.3.10) into the weak form of two-dimensional Poisson problems (Problem 6.1.7) and see how the discretized equation with $\bar{\mathbf{u}}_N$ as an unknown is obtained.

If $u_h(\bar{\mathbf{u}})$ and $v_h(\bar{\mathbf{v}})$ of Eqs. (6.3.1) and (6.3.2) are substituted into the weak form, we get

$$a(u_h(\bar{\mathbf{u}}), v_h(\bar{\mathbf{v}})) = l(v_h(\bar{\mathbf{v}})). \quad (6.3.11)$$

The left-hand side of Eq. (6.3.11) can be written as

$$\begin{aligned} a(u_h(\bar{\mathbf{u}}), v_h(\bar{\mathbf{v}})) &= \sum_{i \in \mathcal{E}} \int_{\Omega_i} \nabla u_h(\bar{\mathbf{u}}_i) \cdot \nabla v_h(\bar{\mathbf{v}}_i) \, dx \\ &= \sum_{i \in \mathcal{E}} a_i(u_h(\bar{\mathbf{u}}_i), v_h(\bar{\mathbf{v}}_i)). \end{aligned} \quad (6.3.12)$$

Here, each term on the right-hand side of Eq. (6.3.12) can be summarized as

$$\begin{aligned} a_i(u_h(\bar{\mathbf{u}}_i), v_h(\bar{\mathbf{v}}_i)) &= (v_{i(1)} \ v_{i(2)} \ v_{i(3)}) \\ &\quad \times \begin{pmatrix} \int_{\Omega_i} \nabla \varphi_{i(1)} \cdot \nabla \varphi_{i(1)} \, dx & \cdots & \int_{\Omega_i} \nabla \varphi_{i(1)} \cdot \nabla \varphi_{i(3)} \, dx \\ \vdots & \ddots & \vdots \\ \int_{\Omega_i} \nabla \varphi_{i(3)} \cdot \nabla \varphi_{i(1)} \, dx & \cdots & \int_{\Omega_i} \nabla \varphi_{i(3)} \cdot \nabla \varphi_{i(3)} \, dx \end{pmatrix} \begin{pmatrix} u_{i(1)} \\ u_{i(2)} \\ u_{i(3)} \end{pmatrix} \\ &= (v_{i(1)} \ v_{i(2)} \ v_{i(3)}) \\ &\quad \times \begin{pmatrix} a_i(\varphi_{i(1)}, \varphi_{i(1)}) & \cdots & a_i(\varphi_{i(1)}, \varphi_{i(3)}) \\ \vdots & \ddots & \vdots \\ a_i(\varphi_{i(3)}, \varphi_{i(1)}) & \cdots & a_i(\varphi_{i(3)}, \varphi_{i(3)}) \end{pmatrix} \begin{pmatrix} u_{i(1)} \\ u_{i(2)} \\ u_{i(3)} \end{pmatrix} \\ &= \bar{\mathbf{v}}_i \cdot \bar{\mathbf{A}}_i \bar{\mathbf{u}}_i = \bar{\mathbf{v}} \cdot \mathbf{Z}_i^\top \bar{\mathbf{A}}_i \mathbf{Z}_i \bar{\mathbf{u}} = \bar{\mathbf{v}} \cdot \tilde{\mathbf{A}}_i \bar{\mathbf{u}}. \end{aligned} \quad (6.3.13)$$

$\tilde{\mathbf{A}}_i$ is called the coefficient matrix of the finite element $i \in \mathcal{E}$. $\tilde{\mathbf{A}}_i$ is a matrix which has been expanded with the addition of 0 to match the total node value vectors.

If η_α and θ_α of Eq. (6.3.7) are substituted into Eq. (6.3.6), $\bar{\mathbf{A}}_i = (\bar{a}_{i(\alpha\beta)})_{\alpha\beta} \in \mathbb{R}^{3 \times 3}$ is calculated as

$$\begin{aligned} \bar{a}_{i(\alpha\beta)} &= \int_{\Omega_i} \left(\frac{\partial \varphi_{i(\alpha)}}{\partial x_1} \frac{\partial \varphi_{i(\beta)}}{\partial x_1} + \frac{\partial \varphi_{i(\alpha)}}{\partial x_2} \frac{\partial \varphi_{i(\beta)}}{\partial x_2} \right) \, dx \\ &= \int_{\Omega_i} (\eta_\alpha \eta_\beta + \theta_\alpha \theta_\beta) \, dx = |\Omega_i| (\eta_\alpha \eta_\beta + \theta_\alpha \theta_\beta), \end{aligned} \quad (6.3.14)$$

where $|\Omega_i| = \gamma/2$ and γ is given by Eq. (6.3.8).

On the other hand, the right-hand side of Eq. (6.3.11) can also be split into elements as

$$\begin{aligned} l(v_h(\bar{\mathbf{v}})) &= \sum_{i \in \mathcal{E}} \int_{\Omega_i} b v_h(\bar{\mathbf{v}}_i) \, dx + \sum_{i \in \mathcal{E}} \int_{\partial \Omega_i \cap \Gamma_N} p_N v_h(\bar{\mathbf{v}}_i) \, dy \\ &= \sum_{i \in \mathcal{E}} l_i(v_h(\bar{\mathbf{v}}_i)), \end{aligned} \quad (6.3.15)$$

where

$$\begin{aligned} l_i(v_h(\bar{\mathbf{v}}_i)) &= (v_{i(1)} \ v_{i(2)} \ v_{i(3)}) \left(\begin{pmatrix} \int_{\Omega_i} b \varphi_{i(1)} \, dx \\ \int_{\Omega_i} b \varphi_{i(2)} \, dx \\ \int_{\Omega_i} b \varphi_{i(3)} \, dx \end{pmatrix} + \begin{pmatrix} \int_{\partial \Omega_i \cap \Gamma_N} p_N \varphi_{i(1)} \, dy \\ \int_{\partial \Omega_i \cap \Gamma_N} p_N \varphi_{i(2)} \, dy \\ \int_{\partial \Omega_i \cap \Gamma_N} p_N \varphi_{i(3)} \, dy \end{pmatrix} \right) \\ &= (v_{i(1)} \ v_{i(2)} \ v_{i(3)}) \left(\begin{pmatrix} b_{i(1)} \\ b_{i(2)} \\ b_{i(3)} \end{pmatrix} + \begin{pmatrix} p_{i(1)} \\ p_{i(2)} \\ p_{i(3)} \end{pmatrix} \right) \\ &= \bar{\mathbf{v}}_i \cdot (\bar{\mathbf{b}}_i + \bar{\mathbf{p}}_i) = \bar{\mathbf{v}} \cdot \left\{ \mathbf{Z}_i^\top (\bar{\mathbf{b}}_i + \bar{\mathbf{p}}_i) \right\} = \bar{\mathbf{v}} \cdot (\tilde{\mathbf{b}}_i + \tilde{\mathbf{p}}_i) \\ &= \bar{\mathbf{v}} \cdot (\mathbf{Z}^\top \tilde{\mathbf{l}}_i) = \bar{\mathbf{v}} \cdot \tilde{\mathbf{l}}_i \end{aligned} \quad (6.3.16)$$

with respect to $i \in \mathcal{E}$. Here, $\tilde{\mathbf{l}}_i$ is called a known term vector of the finite element $i \in \mathcal{E}$. $\bar{\mathbf{b}}_i$ and $\bar{\mathbf{p}}_i$ represent components of the known term vectors with respect to b and p_N respectively. $\tilde{\mathbf{l}}_i$, $\bar{\mathbf{b}}_i$ and $\tilde{\mathbf{p}}_i$ are vectors of $\bar{\mathbf{l}}_i$, $\bar{\mathbf{b}}_i$ and $\bar{\mathbf{p}}_i$ respectively which have been expanded with 0 added in to match the total node value vectors.

If b is a constant function, $\bar{\mathbf{b}}_i = (b_{i(1)}, b_{i(2)}, b_{i(3)})^\top$ is calculated as

$$\bar{\mathbf{b}}_i = b \begin{pmatrix} \int_{\Omega_i} \varphi_{i(1)} \, dx \\ \int_{\Omega_i} \varphi_{i(2)} \, dx \\ \int_{\Omega_i} \varphi_{i(3)} \, dx \end{pmatrix} = \frac{b |\Omega_i|}{3} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}. \quad (6.3.17)$$

Here, the following integration formula using area coordinates was used. Area coordinates are coordinates where a point $\mathbf{x} \in \Omega_i$ in the triangular finite element is represented by the three-dimensional vector with elements of basis functions

$\varphi_{i(1)}(\mathbf{x})$, $\varphi_{i(2)}(\mathbf{x})$ and $\varphi_{i(3)}(\mathbf{x})$ of the triangular finite element (see Sect. 6.4.2 for reference).

Theorem 6.3.1 (Integrals of Area Coordinates) *When $(\varphi_{i(1)}, \varphi_{i(2)}, \varphi_{i(3)})$ denotes the area coordinates on the two-dimensional triangular domain Ω_i ,*

$$\int_{\Omega_i} (\varphi_{i(1)})^l (\varphi_{i(2)})^m (\varphi_{i(3)})^n \, d\mathbf{x} = 2 |\Omega_i| \frac{l!m!n!}{(l+m+n+2)!}$$

holds with respect to non-negative integers l , m and n . Here, $|\Omega_i| = \gamma/2$ and γ is given by Eq. (6.3.8). \square

Furthermore, when p_N is a constant function, $\bar{\mathbf{p}}_i = (p_{i(1)}, p_{i(2)}, p_{i(3)})^\top$ becomes

$$\bar{\mathbf{p}}_i = p_N \begin{pmatrix} 0 \\ \int_{\partial\Omega_i \cap \Gamma_N} \varphi_{i(2)} \, d\gamma \\ \int_{\partial\Omega_i \cap \Gamma_N} \varphi_{i(3)} \, d\gamma \end{pmatrix} = \frac{p_N h}{2} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix},$$

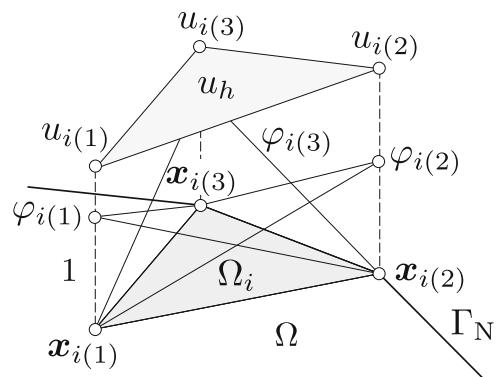
where h is the length of $\partial\Omega_i \cap \Gamma_N$ (see Fig. 6.15).

Here, if Eq. (6.3.12) with Eq. (6.3.13) substituted in and Eq. (6.3.15) with Eq. (6.3.16) substituted in are substituted into the weak form (Eq. (6.3.11)), we get

$$\bar{\mathbf{v}} \cdot \left(\sum_{i \in \mathcal{E}} \tilde{\mathbf{A}}_i \bar{\mathbf{u}} \right) = \bar{\mathbf{v}} \cdot \sum_{i \in \mathcal{E}} \tilde{\mathbf{l}}_i.$$

This equation is written as

Fig. 6.15 Finite element including the boundary



$$\bar{\mathbf{v}} \cdot (\bar{\mathbf{A}} \bar{\mathbf{u}}) = \bar{\mathbf{v}} \cdot \bar{\mathbf{l}}. \quad (6.3.18)$$

In other words,

$$\begin{aligned}\bar{\mathbf{A}} &= \sum_{i \in \mathcal{E}} \tilde{\mathbf{A}}_i \in \mathbb{R}^{|\mathcal{N}| \times |\mathcal{N}|}, \\ \bar{\mathbf{l}} &= \sum_{i \in \mathcal{E}} \tilde{\mathbf{l}}_i = \sum_{i \in \mathcal{E}} (\tilde{\mathbf{b}}_i + \tilde{\mathbf{p}}_i) = \bar{\mathbf{b}} + \bar{\mathbf{p}} \in \mathbb{R}^{|\mathcal{N}|}.\end{aligned}$$

$\bar{\mathbf{A}}$ and $\bar{\mathbf{l}}$ are called the total coefficient matrix and total known term vector, respectively. Moreover, $\bar{\mathbf{b}}$ and $\bar{\mathbf{p}}$ are called total nodal value vectors of \mathbf{b} and \mathbf{p}_N respectively.

We substitute the fundamental boundary conditions into Eq. (6.3.18). In other words, if $u_j = u_D(\mathbf{x}_j)$ and $v_j = 0$ are substituted in $j \in \mathcal{N}_D$,

$$(\mathbf{0}^\top \bar{\mathbf{v}}_N^\top) \left(\begin{pmatrix} \bar{\mathbf{A}}_{DD} & \bar{\mathbf{A}}_{DN} \\ \bar{\mathbf{A}}_{ND} & \bar{\mathbf{A}}_{NN} \end{pmatrix} \begin{pmatrix} \bar{\mathbf{u}}_D \\ \bar{\mathbf{u}}_N \end{pmatrix} - \begin{pmatrix} \bar{\mathbf{l}}_D \\ \bar{\mathbf{l}}_N \end{pmatrix} \right) = 0 \quad (6.3.19)$$

is obtained. Here, $\bar{\mathbf{u}}_D$ and $\bar{\mathbf{u}}_N$ are vectors defined in Eq. (6.3.1) and $\bar{\mathbf{v}}_D$ and $\bar{\mathbf{v}}_N$ are vectors defined in Eq. (6.3.2). Moreover,

$$\begin{aligned}\bar{\mathbf{A}}_{DD} &= (\bar{A}_{ij})_{i \in \mathcal{N}_D, j \in \mathcal{N}_D}, & \bar{\mathbf{A}}_{DN} &= (\bar{A}_{ij})_{i \in \mathcal{N}_D, j \in \mathcal{N}_N}, \\ \bar{\mathbf{A}}_{ND} &= (\bar{A}_{ij})_{i \in \mathcal{N}_N, j \in \mathcal{N}_D}, & \bar{\mathbf{A}}_{NN} &= (\bar{A}_{ij})_{i \in \mathcal{N}_N, j \in \mathcal{N}_N}, \\ \bar{\mathbf{l}}_D &= (l_i)_{i \in \mathcal{N}_D}, & \bar{\mathbf{l}}_N &= (l_i)_{i \in \mathcal{N}_N}.\end{aligned}$$

Rearranging Eq. (6.3.19) gives

$$\bar{\mathbf{v}}_N^\top (\bar{\mathbf{A}}_{NN} \bar{\mathbf{u}}_N - \bar{\mathbf{l}}_N + \bar{\mathbf{u}}_D \bar{\mathbf{A}}_{ND}) = 0.$$

Since $\bar{\mathbf{v}}_N$ is arbitrary, we can write

$$\bar{\mathbf{A}}_{NN} \bar{\mathbf{u}}_N = \bar{\mathbf{l}}_N - \bar{\mathbf{u}}_D \bar{\mathbf{A}}_{ND} = \hat{\mathbf{l}}. \quad (6.3.20)$$

Equation (6.3.20) is a simultaneous linear equation with respect to the unknown vector $\bar{\mathbf{u}}_N$ and is called the discretized equation of the finite element method. If Eq. (6.3.20) is solved for $\bar{\mathbf{u}}_N$, we get

$$\bar{\mathbf{u}}_N = \bar{\mathbf{A}}_{NN}^{-1} \hat{\mathbf{l}}. \quad (6.3.21)$$

From these, the finite element solution $u_h(\bar{\mathbf{u}})$ can be obtained by substituting $\bar{\mathbf{u}} = (\bar{\mathbf{u}}_D, \bar{\mathbf{u}}_N^\top)^\top$ into Eq. (6.3.1). Moreover, the finite element solution $u_h(\bar{\mathbf{u}}_i)$ in the domain Ω_i of the finite element $i \in \mathcal{E}$ can be sought via Eq. (6.3.9).

6.3.4 Exercise Problem

Let us look in detail at how an approximate solution of a two-dimensional Poisson problem can be obtained using the triangular finite element shown above [96, Section 5.3, p. 67].

Exercise 6.3.2 (Finite Element Method for 2D Poisson Problem) Let the domain Ω be $(0, 1)^2$ and define the boundaries $\Gamma_D = \{x \in \partial\Omega \mid x_1 = 0, x_2 = 0\}$ and $\Gamma_N = \partial\Omega \setminus \Gamma_D$. In this case, obtain the approximate solution from the finite element method of $u : (0, 1)^2 \rightarrow \mathbb{R}$ which satisfies

$$-\Delta u = 1 \quad \text{in } \Omega, \quad \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \Gamma_N, \quad u = 0 \quad \text{on } \Gamma_D.$$

Here, use the element partition in Fig. 6.16. □

Answer Let the size of the finite element be $h = 1/2$. In this case, from Eq. (6.3.8), we get $\gamma = h^2$ and $|\Omega_i| = \gamma/2 = h^2/2$. Let us calculate the coefficient matrix \bar{A}_i and \bar{b}_i using Eqs. (6.3.14) and (6.3.17). Think about the finite element split into two types. Let the finite element $i \in \{1, 3, 5, 7\}$ of the form such as in Fig. 6.17a be as type 1. The basis function $\varphi_{i(\alpha)}$ with respect to $\alpha \in \{1, 2, 3\}$ of type 1 was given by Eq. (6.3.6). The undetermined multipliers which are required in the calculation of

Fig. 6.16 An example of a 2D Poisson problem. (a) Domain Ω . (b) Finite element partition

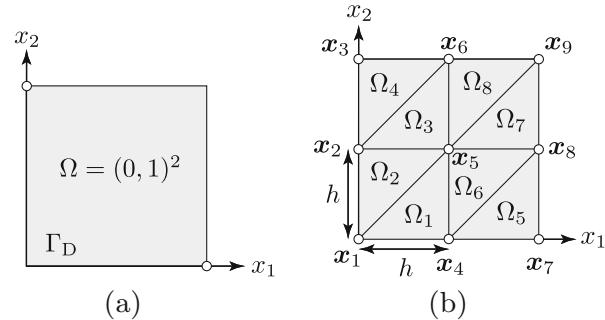
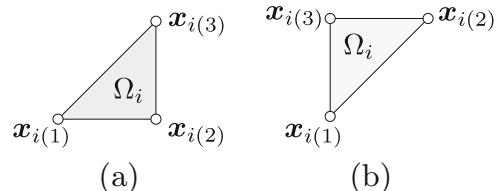


Fig. 6.17 Finite element types. (a) Type 1. (b) Type 2



the coefficient matrix \bar{A}_i are η_α and θ_α . These values with respect to type 1 are

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix} = \frac{1}{\gamma} \begin{pmatrix} x_{i(2)2} - x_{i(3)2} \\ x_{i(3)2} - x_{i(1)2} \\ x_{i(1)2} - x_{i(2)2} \end{pmatrix} = \frac{1}{h^2} \begin{pmatrix} -h \\ h \\ 0 \end{pmatrix},$$

$$\begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} = \frac{1}{\gamma} \begin{pmatrix} x_{i(3)1} - x_{i(2)1} \\ x_{i(1)1} - x_{i(3)1} \\ x_{i(2)1} - x_{i(1)1} \end{pmatrix} = \frac{1}{h^2} \begin{pmatrix} 0 \\ -h \\ h \end{pmatrix}$$

from Eq. (6.3.7). Substituting these into Eqs. (6.3.14) and (6.3.17) gives the coefficient matrix and the known term vector as

$$\bar{A}_i = \frac{1}{2} \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}, \quad \bar{b}_i = \frac{h^2}{6} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

On the other hand, the finite element $i \in \{2, 4, 6, 8\}$ in the form such as the one in Fig. 6.17b is type 2. Similarly, with respect to these, the undetermined multipliers of the basis function can be obtained as

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix} = \frac{1}{\gamma} \begin{pmatrix} x_{i(2)2} - x_{i(3)2} \\ x_{i(3)2} - x_{i(1)2} \\ x_{i(1)2} - x_{i(2)2} \end{pmatrix} = \frac{1}{h^2} \begin{pmatrix} 0 \\ h \\ -h \end{pmatrix},$$

$$\begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} = \frac{1}{\gamma} \begin{pmatrix} x_{i(3)1} - x_{i(2)1} \\ x_{i(1)1} - x_{i(3)1} \\ x_{i(2)1} - x_{i(1)1} \end{pmatrix} = \frac{1}{h^2} \begin{pmatrix} -h \\ 0 \\ h \end{pmatrix}.$$

Substituting these into Eqs. (6.3.14) and (6.3.17) gives a type 2 coefficient matrix and known term vector as

$$\bar{A}_i = \frac{1}{2} \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 2 \end{pmatrix}, \quad \bar{b}_i = \frac{h^2}{6} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

The relationships of \mathbf{x}_j with respect to the local nodes $\mathbf{x}_{i(1)}$, $\mathbf{x}_{i(2)}$, $\mathbf{x}_{i(3)}$ and total nodes $j \in \mathcal{N} = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ for the finite element $i \in \mathcal{E} = \{1, 2, 3, 4, 5, 6, 7, 8\}$ are shown in Table 6.1.

Table 6.1 Relationship between local nodes $\mathbf{x}_{i(1)}$, $\mathbf{x}_{i(2)}$, $\mathbf{x}_{i(3)}$ and the total nodes \mathbf{x}_j in Exercise 6.3.2

$i \in \mathcal{E}$	1	2	3	4	5	6	7	8
$\mathbf{x}_{i(1)}$	\mathbf{x}_1	\mathbf{x}_1	\mathbf{x}_2	\mathbf{x}_2	\mathbf{x}_4	\mathbf{x}_4	\mathbf{x}_5	\mathbf{x}_5
$\mathbf{x}_{i(2)}$	\mathbf{x}_4	\mathbf{x}_5	\mathbf{x}_5	\mathbf{x}_6	\mathbf{x}_7	\mathbf{x}_8	\mathbf{x}_8	\mathbf{x}_9
$\mathbf{x}_{i(3)}$	\mathbf{x}_5	\mathbf{x}_2	\mathbf{x}_6	\mathbf{x}_3	\mathbf{x}_8	\mathbf{x}_5	\mathbf{x}_9	\mathbf{x}_6
Type	1	2	1	2	1	2	1	2

We expand $\tilde{\mathbf{A}}_1$ and $\tilde{\mathbf{l}}_1$ in conjunction with the total node value vectors as

$$\tilde{\mathbf{A}}_1 = \frac{1}{2} \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \tilde{\mathbf{l}}_1 = \frac{h^2}{6} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Similarly, if $\tilde{\mathbf{A}}_i$ and $\tilde{\mathbf{l}}_i$ are formed using $\bar{\mathbf{A}}_i$ and $\bar{\mathbf{l}}_i$ with respect to $i \in \{2, \dots, 8\}$ and superimposed, one obtains

$$\tilde{\mathbf{A}} = \frac{1}{2} \begin{pmatrix} 2 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -2 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 4 & -2 & 0 & -1 & 0 & 0 \\ 0 & -2 & 0 & -2 & 8 & -2 & 0 & -2 & 0 \\ 0 & 0 & -1 & 0 & -2 & 4 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & -2 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 2 \end{pmatrix}, \quad \tilde{\mathbf{l}} = \frac{h^2}{6} \begin{pmatrix} 2 \\ 3 \\ 1 \\ 3 \\ 6 \\ 3 \\ 1 \\ 3 \\ 2 \end{pmatrix}.$$

Moreover, if the fundamental boundary conditions $u_1 = u_2 = u_3 = u_4 = u_7 = 0$, $v_1 = v_2 = v_3 = v_4 = v_7 = 0$ are substituted in as Eq. (6.3.19), we get

$$\frac{1}{2} \begin{pmatrix} 8 & -2 & -2 & 0 \\ -2 & 4 & 0 & -1 \\ -2 & 0 & 4 & -1 \\ 0 & -1 & -1 & 2 \end{pmatrix} \begin{pmatrix} u_5 \\ u_6 \\ u_8 \\ u_9 \end{pmatrix} = \frac{1}{24} \begin{pmatrix} 6 \\ 3 \\ 3 \\ 2 \end{pmatrix}.$$

Solving this simultaneous linear equation,

$$\begin{pmatrix} u_5 \\ u_6 \\ u_8 \\ u_9 \end{pmatrix} = \frac{1}{192} \begin{pmatrix} 3 & 2 & 2 & 2 \\ 2 & 6 & 2 & 4 \\ 2 & 2 & 6 & 4 \\ 2 & 4 & 4 & 12 \end{pmatrix} \begin{pmatrix} 6 \\ 3 \\ 3 \\ 2 \end{pmatrix} = \frac{1}{96} \begin{pmatrix} 17 \\ 22 \\ 22 \\ 30 \end{pmatrix}$$

is obtained. Figure 6.18 shows the approximate solution u_h (\bar{u}) with this result as a node value. \square

Figure 6.19 shows the result when the division number in Exercise 6.3.2 is increased to 36 nodes and 50 finite elements.

Fig. 6.18 Approximate solution of Exercise 6.3.2

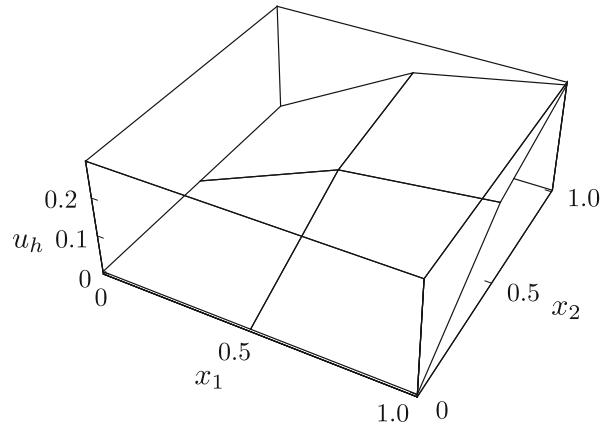
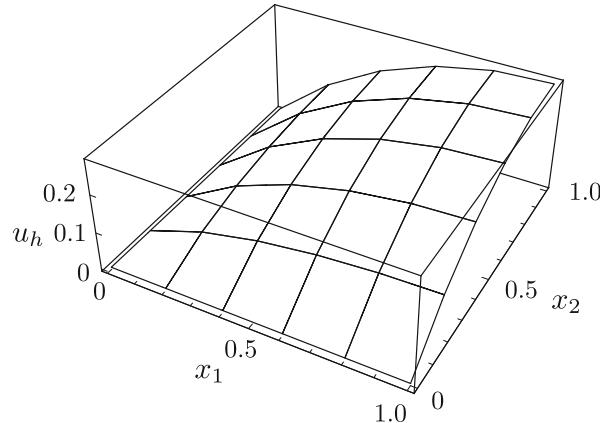


Fig. 6.19 Approximate solution when increasing the division number of Exercise 6.3.2



6.4 Various Finite Elements

The finite element methods in one and two dimensions when the basis function is constructed of linear functions were seen in Sects. 6.2 and 6.3. Next, let us show how to change the basis function to a higher order and how to change the shape of a finite element into a square, and give an overview of the three-dimensional finite element method. Other than the method of construction of the basis function being different, their procedures, such as the Galerkin method seeking the undetermined multipliers by substituting them into the weak form, are identical. So let us look at how the basis function is defined for a finite element and how the approximate function is constructed.

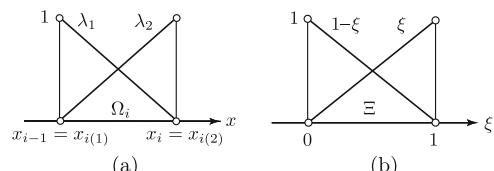
In this chapter, the domain of the basis function is changed to a domain whose size is normalized (normal domain), and a finite element defined on such a domain is referred to as normal element. A finite element of arbitrary size is thought to be given by a mapping from a normal element.

6.4.1 One-Dimensional Higher-Order Finite Elements

Firstly, let us think about making the one-dimensional finite element a higher-order. Let the normal domain with respect to the one-dimensional finite element be $\Xi = (0, 1)$. The point $x \in \Omega_i$ on the domain $\Omega_i = (x_{i-1}, x_i)$ of the one-dimensional finite element $i \in \mathcal{E}$ can be changed to be a point on the normal domain $\xi \in \Xi$ by using the basis functions $\varphi_{i(1)}$ and $\varphi_{i(2)}$ of the one-dimensional finite element (first-order element) defined by Eqs. (6.2.3) and (6.2.4) via $\xi = \varphi_{i(2)}(x) = 1 - \varphi_{i(1)}(x)$. Here, $\varphi_{i(2)}(x) = 1 - \varphi_{i(1)}(x)$ and $\xi \in \Xi$ are viewed as the same and defined $(\varphi_{i(1)}(x), \varphi_{i(2)}(x))$ as the length coordinate. It is a two-dimensional vector to express length, but it should be noted that the condition $\varphi_{i(1)}(x) + \varphi_{i(2)}(x) = 1$ is imposed. Furthermore, using functions of $(\varphi_{i(1)}(x), \varphi_{i(2)}(x))$ as coordinates may bring confusion so the length coordinate is written as (λ_1, λ_2) . Figure 6.20 shows the relationship between the length coordinates (λ_1, λ_2) and normal coordinates ξ .

Hereafter, we consider the basis functions defined on the normal domain $\Xi = (0, 1)$ and the basis functions will be made higher-order. As a preparation for this, let us confirm how the basis functions $(\varphi_{i(1)}(x), \varphi_{i(2)}(x))$ defined with respect to $x \in \Omega_i$ seen in Sect. 6.2 is expressed using the basis functions defined with respect to $\xi \in \Xi$. In this chapter, the node and basis function defined on the normal domain

Fig. 6.20 Length coordinates and normal coordinates of 1D finite elements. (a) Length coordinates (λ_1, λ_2) . (b) Normal coordinates $\xi = \lambda_2 = 1 - \lambda_1$



are to be expressed as $\xi(\cdot)$ and $\hat{\phi}(\cdot)$, respectively. In other words, let the node with respect to first-order basis function be

$$(\xi_{(1)} \ \xi_{(2)}) = (0 \ 1) \quad (6.4.1)$$

and the basis function be

$$(\hat{\phi}_{(1)}(\xi) \ \hat{\phi}_{(2)}(\xi)) = (\lambda_1 \ \lambda_2) = (1 - \xi \ \xi). \quad (6.4.2)$$

On the other hand, the mapping $f_i : \Xi \rightarrow \Omega_i$ from normal coordinate $\xi \in \Xi$ to the global coordinate $x \in \Omega_i$ is given by

$$x = f_i(\xi) = x_{i(1)} + \xi (x_{i(2)} - x_{i(1)}). \quad (6.4.3)$$

Here, between the first-order basis functions $(\varphi_{i(1)}, \varphi_{i(2)})$ defined on Ω_i and the first-order basis functions $(\hat{\phi}_{(1)}, \hat{\phi}_{(2)})$ defined on the normal domain,

$$(\varphi_{i(1)}(f_i(\xi)) \ \varphi_{i(2)}(f_i(\xi))) = (\hat{\phi}_{(1)}(\xi) \ \hat{\phi}_{(2)}(\xi)) \quad (6.4.4)$$

holds.

Let us define second-order basis functions corresponding to the definition of first-order basis functions. Second-order functions have three undetermined multipliers. Hence, adding a mid-node, we define the nodes as

$$(\xi_{i(1)} \ \xi_{i(2)} \ \xi_{i(3)}) = (0 \ 1 \ 1/2). \quad (6.4.5)$$

The basis functions $\hat{\phi}_{(1)}$, $\hat{\phi}_{(2)}$ and $\hat{\phi}_{(3)}$ of normal element are determined so that they satisfy the boundary conditions at $\xi_{(1)}$, $\xi_{(2)}$ and $\xi_{(3)}$, respectively. These conditions can be given by

$$\hat{\phi}_{(\alpha)}(\xi_{(\beta)}) = \delta_{\alpha\beta} \quad (6.4.6)$$

with respect to $\alpha, \beta \in \{1, 2, 3\}$. Equation (6.4.6) shows conditions for the undetermined multipliers to be the node values of an approximate function. From this condition, the three undetermined multipliers of a second-order function are determined as

$$(\hat{\phi}_{(1)} \ \hat{\phi}_{(2)} \ \hat{\phi}_{(3)}) = (\lambda_1 (2\lambda_1 - 1) \ \lambda_2 (2\lambda_2 - 1) \ 4\lambda_1\lambda_2). \quad (6.4.7)$$

Basis functions determined in this way satisfies

$$\sum_{\alpha \in \{1, 2, 3\}} \hat{\phi}_{(\alpha)}(\xi) = 1 \quad (6.4.8)$$

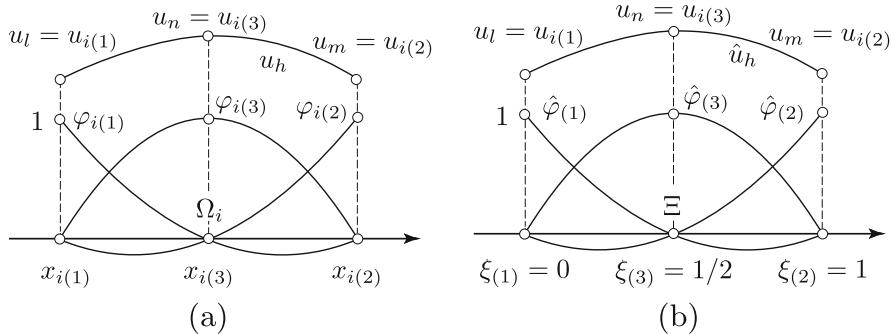


Fig. 6.21 Basis functions and an approximate function used in a second-order 1D finite element.
(a) Basis functions on Ω_i . **(b)** Basis functions on Ξ

with respect to all $\xi \in \Xi$. Equation (6.4.8) is the condition at which the approximate solution matches the exact solution when the exact solution is $u = 1$. Figure 6.21 shows these basis functions and the basis functions defined on Ω_i . Here, let l, m , and n in the figure be the node numbers given in total nodes.

Therefore, the approximate function used in a second-order one-dimensional finite element is constructed as

$$\hat{u}_h(\xi) = (\hat{\varphi}_{(1)}(\xi) \ \hat{\varphi}_{(2)}(\xi) \ \hat{\varphi}_{(3)}(\xi)) \begin{pmatrix} u_{i(1)} \\ u_{i(2)} \\ u_{i(3)} \end{pmatrix} = \hat{\varphi}(\xi) \cdot \bar{u}_i$$

on the normal domain. Here, in this chapter the approximate function defined on a normal domain will be expressed as \hat{u}_h .

Similarly, the one-dimensional finite element of $m (\in \mathbb{N})$ -th order is constructed in the following way.

Definition 6.4.1 (m -th Order One-Dimensional Finite Element) Let $\Xi = (0, 1)$ be a normal domain. With respect to $m \in \mathbb{N}$, place the nodes as

$$(\xi_{i(1)} \ \xi_{i(2)} \ \cdots \ \xi_{i(m+1)}) = (0 \ 1/m \ \cdots \ 1).$$

Construct the basis functions $\hat{\varphi}_{(\alpha)}$ with respect to $\alpha \in \{1, \dots, m+1\}$ as an m -th order polynomial and select the undetermined multipliers so that Eq. (6.4.6) with respect to $\beta \in \{1, \dots, m+1\}$ is satisfied. A finite element using basis functions constructed in this way is called a one-dimensional m -th order finite element. \square

When basis functions $\hat{\varphi}_{(\alpha)}$ are defined as in Definition 6.4.1, an approximate function is constructed by

$$\hat{u}_h = \sum_{\alpha \in \mathcal{N}_i} \hat{\varphi}_{(\alpha)} u_{i(\alpha)} = \hat{\varphi} \cdot \bar{u}_i \quad (6.4.9)$$

on the normal domain, where \mathcal{N}_i is a set of local node numbers. Equation (6.4.9) expresses the relationship which holds with respect to the approximate function, not limited to an m -th order one-dimensional finite element.

In this way, if approximate functions are given on the normal coordinates $\xi \in \Xi$, the bilinear form for each finite element in the weak form (Eq. (6.2.13)) is

$$\begin{aligned}
 a_i(u_h(\bar{\mathbf{u}}_i), v_h(\bar{\mathbf{v}}_i)) \\
 = (v_{i(1)} \cdots v_{i(m+1)}) \\
 \times \begin{pmatrix} a_i(\varphi_{i(1)}, \varphi_{i(1)}) & \cdots & a_i(\varphi_{i(1)}, \varphi_{i(m+1)}) \\ \vdots & \ddots & \vdots \\ a_i(\varphi_{i(m+1)}, \varphi_{i(1)}) & \cdots & a_i(\varphi_{i(m+1)}, \varphi_{i(m+1)}) \end{pmatrix} \begin{pmatrix} u_{i(1)} \\ \vdots \\ u_{i(m+1)} \end{pmatrix} \\
 = \bar{\mathbf{v}}_i \cdot (\bar{\mathbf{A}}_i \bar{\mathbf{u}}_i) = \bar{\mathbf{v}} \cdot (\mathbf{Z}_i^\top \bar{\mathbf{A}}_i \mathbf{Z}_i \bar{\mathbf{u}}) = \bar{\mathbf{v}} \cdot (\tilde{\mathbf{A}}_i \bar{\mathbf{u}}). \tag{6.4.10}
 \end{aligned}$$

Here, we can write

$$\frac{d\varphi_{i(\alpha)}}{dx} = \frac{d\hat{\varphi}_{(\alpha)}}{d\xi} \frac{dx}{d\xi} = \frac{1}{\omega_i} \frac{d\hat{\varphi}_{(\alpha)}}{d\xi},$$

where

$$\omega_i = \frac{dx}{d\xi} = \frac{df_i}{d\xi} = x_{i(2)} - x_{i(1)}$$

from Eq. (6.4.3). Hence, each element of $\bar{\mathbf{A}}_i = (\bar{a}_{i(\alpha\beta)})_{\alpha\beta} \in \mathbb{R}^{|\mathcal{N}_i| \times |\mathcal{N}_i|}$ can be calculated from

$$\bar{a}_{i(\alpha\beta)} = \int_{\Omega_i} \frac{d\varphi_{i(\alpha)}}{dx} \frac{d\varphi_{i(\beta)}}{dx} dx = \frac{1}{\omega_i} \int_0^1 \frac{d\hat{\varphi}_{(\alpha)}}{d\xi} \frac{d\hat{\varphi}_{(\beta)}}{d\xi} d\xi.$$

Moreover, even with respect to the linear form (Eq. (6.2.16)) for each finite element in the weak form, we have

$$\begin{aligned}
 l_i(v_h(\bar{\mathbf{v}}_i)) &= (v_{i(1)} \cdots v_{i(m+1)}) \begin{pmatrix} b_{i(1)} \\ \vdots \\ b_{i(m+1)} \end{pmatrix} \\
 &= \bar{\mathbf{v}}_i \cdot \bar{\mathbf{b}}_i = \bar{\mathbf{v}} \cdot (\mathbf{Z}_i^\top \bar{\mathbf{b}}_i) = \bar{\mathbf{v}} \cdot \tilde{\mathbf{b}}_i \\
 &= \bar{\mathbf{v}}_i \cdot \bar{\mathbf{l}}_i = \bar{\mathbf{v}} \cdot (\mathbf{Z}_i^\top \bar{\mathbf{l}}_i) = \bar{\mathbf{v}} \cdot \tilde{\mathbf{l}}_i. \tag{6.4.11}
 \end{aligned}$$

Here, each element of $\bar{\mathbf{b}}_i = (\bar{b}_{i(\alpha)})_\alpha \in \mathbb{R}^{|\mathcal{N}_i|}$ can be calculated by

$$\bar{b}_{i(\alpha)} = \int_{\Omega_i} b_0 \varphi_{i(\alpha)} \, dx = \omega_i \int_0^1 b_0 \hat{\varphi}_{(\alpha)} \, d\xi.$$

A similar relationship holds with respect to Eq. (6.2.17) too.

6.4.2 Triangular Higher-Order Finite Elements

Next, let us think about the higher-order triangular finite element used with respect to a two-dimensional problem. Basis functions $\varphi_{i(1)}$, $\varphi_{i(2)}$ and $\varphi_{i(3)}$ of the first-order triangular finite element defined by Eq. (6.3.6) are called area coordinates. The reason for this is because when a point $\mathbf{x} \in \Omega_i$ on the domain Ω_i of finite element $i \in \mathcal{E}$ is selected as in Fig. 6.22 and λ_1 , λ_2 and λ_3 are the area ratios of three small triangles with \mathbf{x} as a vertex against $|\Omega_i|$, $(\lambda_1, \lambda_2, \lambda_3) = (\varphi_{i(1)}, \varphi_{i(2)}, \varphi_{i(3)})$ is established. From this, $\xi = (\xi_1, \xi_2)^\top = (\lambda_2, \lambda_3)^\top \in \mathbb{R}^2$ is called the normal coordinates and $\Xi = \{ \xi \in (0, 1)^2 \mid \xi_1 + \xi_2 < 1 \}$ is called a normal domain.

Let us again think about the higher-order after defining linear basis functions on the normal domain. Let nodes with respect to normal element be

$$(\hat{\varphi}_{(1)}(\xi) \hat{\varphi}_{(2)}(\xi) \hat{\varphi}_{(3)}(\xi)) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (6.4.12)$$

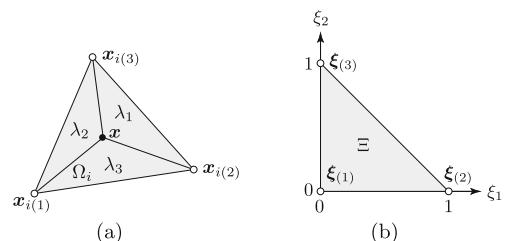
The basis functions looked at in Sect. 6.3 were

$$(\hat{\varphi}_{(1)}(\xi) \hat{\varphi}_{(2)}(\xi) \hat{\varphi}_{(3)}(\xi)) = (\lambda_1 \lambda_2 \lambda_3). \quad (6.4.13)$$

Let us define the second-order basis functions corresponding to the definition of the first-order basis functions. Complete second-order polynomials with respect to ξ_1 and ξ_2 have six undetermined multipliers a_1, \dots, a_6 given as

$$a_1 + a_2 \xi_1 + a_3 \xi_2 + a_4 \xi_1^2 + a_5 \xi_1 \xi_2 + a_6 \xi_2^2.$$

Fig. 6.22 Area coordinates and normal coordinates for a triangular finite element. (a) Area coordinates $(\lambda_1, \lambda_2, \lambda_3)$. (b) Normal coordinates $(\xi_1, \xi_2) = (\lambda_2, \lambda_3)$



Then, mid-point nodes can be added to the three sides of the triangle as shown in Fig. 6.23. Let the nodes be

$$(\xi_{(1)} \ \xi_{(2)} \ \xi_{(3)} \ \xi_{(4)} \ \xi_{(5)} \ \xi_{(6)}) = \begin{pmatrix} 0 & 1 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1 & 1/2 & 1/2 & 0 \end{pmatrix}. \quad (6.4.14)$$

Here, if the basis functions $\hat{\varphi}_{(\alpha)}$ of the normal element with respect to $\alpha \in \{1, \dots, 6\}$ are determined so that $(\xi_{(\beta)}) = \delta_{\alpha\beta}$ is satisfied with respect to $\beta \in \{1, \dots, 6\}$, we get

$$\begin{pmatrix} \hat{\varphi}_{(1)} \\ \hat{\varphi}_{(2)} \\ \hat{\varphi}_{(3)} \end{pmatrix} = \begin{pmatrix} \lambda_1 (2\lambda_1 - 1) \\ \lambda_2 (2\lambda_2 - 1) \\ \lambda_3 (2\lambda_3 - 1) \end{pmatrix}, \quad \begin{pmatrix} \hat{\varphi}_{(4)} \\ \hat{\varphi}_{(5)} \\ \hat{\varphi}_{(6)} \end{pmatrix} = \begin{pmatrix} 4\lambda_2\lambda_3 \\ 4\lambda_1\lambda_3 \\ 4\lambda_1\lambda_2 \end{pmatrix}.$$

Figure 6.24 shows these functions. An approximate function can be constructed using these as Eq. (6.4.9).

Similarly, an $m (\in \mathbb{N})$ -th-order triangular finite element can be constructed in the following way from a two-dimensional m -th order complete polynomial.

Definition 6.4.2 (m-th Order Triangular Finite Element) Let $\Xi = \{\xi \in (0, 1)^2 \mid \xi_1 + \xi_2 < 1\}$ be a normal domain. For $m \in \mathbb{N}$, place the nodes $\xi_{(\alpha)}$ with respect to

Fig. 6.23 Nodes used in second-order triangular finite element. (a) Nodes on Ω_i . (b) Nodes on Ξ

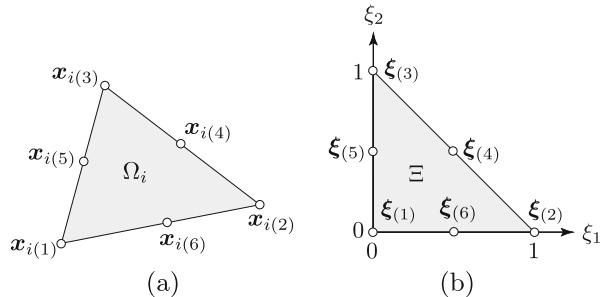


Fig. 6.24 Basis functions and an approximate function used in the second-order triangular finite element

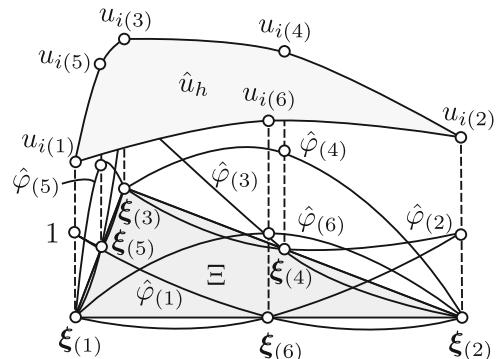
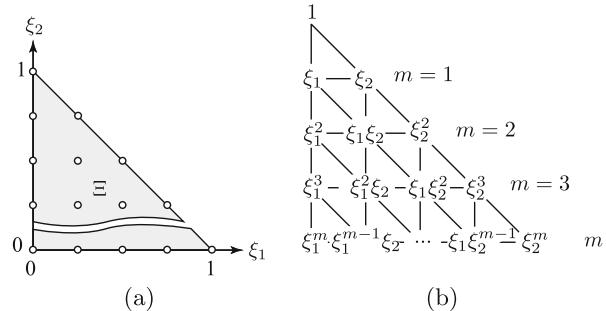


Fig. 6.25 Node placement and polynomial terms used in the m -th order triangular finite element. (a) Nodes $\xi_{(\alpha)}$ for $\alpha \in \mathcal{N}_i$. (b) Complete m -th order polynomial terms of $|\mathcal{N}_i|$



$\alpha \in \mathcal{N}_i$ as shown in Fig. 6.25. Construct the basis functions $\hat{\varphi}_{(\alpha)}$ with a complete m -th order polynomial with respect to ξ_1 and ξ_2 and determine the undetermined multipliers so that $\hat{\varphi}_{(\alpha)}(\xi_{(\beta)}) = \delta_{\alpha\beta}$ is satisfied with respect to $\alpha, \beta \in \mathcal{N}_i$. The finite element using the basis functions constructed in this way is called triangular m -th order finite element. \square

When using the basis functions $\hat{\varphi}_{(\alpha)}$ constructed as in Definition 6.4.2, an approximate function can be constructed on the normal domain as Eq. (6.4.9). If approximate functions are given, the bilinear form (Eq. (6.3.14)) for each finite element in the weak form can be calculated by

$$\begin{aligned} \bar{a}_{i(\alpha\beta)} &= \int_{\Omega_i} \partial_{\mathbf{x}} \varphi_{i(\alpha)}(\mathbf{x}) \cdot \partial_{\mathbf{x}} \varphi_{i(\beta)}(\mathbf{x}) \, d\mathbf{x} \\ &= \int_{\Xi} \left\{ \left(\mathbf{F}_i^{\top} \right)^{-1} \partial_{\xi} \hat{\varphi}_{(\alpha)}(\xi) \right\} \cdot \left\{ \left(\mathbf{F}_i^{\top} \right)^{-1} \partial_{\xi} \hat{\varphi}_{(\beta)}(\xi) \right\} \omega_i \, d\xi. \end{aligned} \quad (6.4.15)$$

Here, the mapping from the normal domain Ξ to the domain Ω_i of the finite element $i \in \mathcal{E}$ was assumed to be given by

$$\begin{aligned} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &= \begin{pmatrix} x_{i(2)1} - x_{i(1)1} & x_{i(3)1} - x_{i(1)1} \\ x_{i(2)2} - x_{i(1)2} & x_{i(3)2} - x_{i(1)2} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} x_{i(1)1} \\ x_{i(1)2} \end{pmatrix} \\ &= \mathbf{F}_i \xi + \mathbf{x}_{i(1)}. \end{aligned} \quad (6.4.16)$$

Here, $\mathbf{x}_{i(\alpha)} = (x_{i(\alpha)1}, x_{i(\alpha)2})^{\top}$ are the coordinate values of local nodes $\alpha \in \{1, 2, 3\}$ of the finite element $i \in \mathcal{E}$. A mapping such as combining a linear mapping and a translation with a fixed element is called affine mapping. In this mapping, \mathbf{F}_i represents a Jacobi matrix and ω_i represents a Jacobian $\det \mathbf{F}_i$. Moreover,

$$\begin{aligned} \partial_{\xi} \hat{\varphi}_{(\alpha)}(\xi) &= \begin{pmatrix} \partial \hat{\varphi}_{(\alpha)} / \partial \xi_1 \\ \partial \hat{\varphi}_{(\alpha)} / \partial \xi_2 \end{pmatrix} = \begin{pmatrix} \partial x_1 / \partial \xi_1 & \partial x_2 / \partial \xi_1 \\ \partial x_1 / \partial \xi_2 & \partial x_2 / \partial \xi_2 \end{pmatrix} \begin{pmatrix} \partial \hat{\varphi}_{(\alpha)} / \partial x_1 \\ \partial \hat{\varphi}_{(\alpha)} / \partial x_2 \end{pmatrix} \\ &= \mathbf{F}_i^{\top} \partial_{\mathbf{x}} \hat{\varphi}_{(\alpha)}(\xi) \end{aligned}$$

was used to get Eq. (6.4.15).

Furthermore, the linear form (Eq. (6.3.16)) for each finite element in the weak form becomes

$$\begin{aligned} l_i(v_h(\bar{v}_i)) &= \bar{v}_i \cdot (\bar{b}_i + \bar{p}_i) = \bar{v} \cdot \left\{ \mathbf{Z}_i^\top (\bar{b}_i + \bar{p}_i) \right\} = \bar{v} \cdot (\tilde{b}_i + \tilde{p}_i) \\ &= \bar{v}_i \cdot \bar{l}_i = \bar{v} \cdot \left(\mathbf{Z}_i^\top \bar{l}_i \right) = \bar{v} \cdot \tilde{l}_i. \end{aligned} \quad (6.4.17)$$

Here, each of the elements in $\bar{b}_i = (\bar{b}_{i(\alpha)})_\alpha \in \mathbb{R}^{|\mathcal{N}_i|}$ and $\bar{p}_i = (\bar{p}_{i(\alpha)})_\alpha \in \mathbb{R}^{|\mathcal{N}_i|}$ are calculated from

$$\bar{b}_{i(\alpha)} = \int_{\Omega_i} b_0 \varphi_{i(\alpha)} \, dx = \omega_i \int_{\Xi} b_0 \hat{\varphi}_{(\alpha)} \, d\xi, \quad (6.4.18)$$

$$\bar{p}_{i(\alpha)} = \int_{\partial\Omega_i \cap \Gamma_N} \varphi_{i(2)} \, d\gamma = \omega_{i1D} \int_0^1 p_N \hat{\varphi}_{(\alpha)} \, d\xi_1, \quad (6.4.19)$$

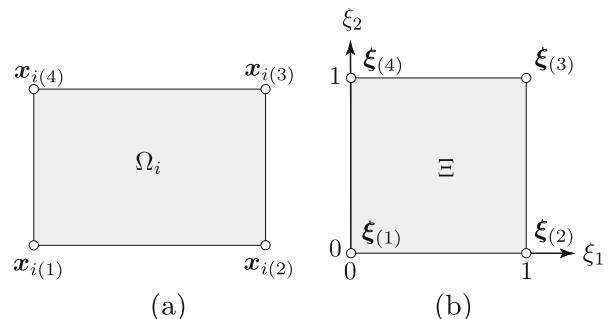
where $\omega_{i1D} = d\gamma/d\xi_1 = |\partial\Omega_i \cap \Gamma_N|$.

The integral on the triangular normal domain Ξ in the above equation can be calculated using the formula in Theorem 6.3.1.

6.4.3 Rectangular Finite Elements

A rectangular finite element can also be considered with respect to two-dimensional problems. Consider two-dimensional domain Ω that can be divided with rectangular domains Ω_i ($i \in \mathcal{E}$) such as the one in Fig. 6.26a. Let the normal domain Ξ be $(0, 1)^2$ such as the one in Fig. 6.26b. The change from $x \in \Omega_i$ to $\xi \in \Xi$ is given by $\xi = (\xi_1, \xi_2)^\top = (\lambda_{12}, \lambda_{22})^\top \in \Xi$ based on length coordinate in the x_1 direction

Fig. 6.26 Node of a rectangular finite element. (a) Node on Ω_i . (b) Node on Ξ



$(\lambda_{11}, \lambda_{12})$ and length coordinates in the x_2 direction $(\lambda_{21}, \lambda_{22})$ which are defined as

$$\lambda_{11}(\mathbf{x}) = \frac{x_{i(2)1} - x_{11}}{x_{i(2)1} - x_{i(1)1}}, \quad (6.4.20)$$

$$\lambda_{12}(\mathbf{x}) = \frac{x_{11} - x_{i(1)1}}{x_{i(2)1} - x_{i(1)1}}, \quad (6.4.21)$$

$$\lambda_{21}(\mathbf{x}) = \frac{x_{i(4)2} - x_{21}}{x_{i(4)2} - x_{i(1)2}}, \quad (6.4.22)$$

$$\lambda_{22}(\mathbf{x}) = \frac{x_{21} - x_{i(1)2}}{x_{i(4)2} - x_{i(1)2}}. \quad (6.4.23)$$

In a first-order rectangular finite element, the bilinear polynomial

$$a_1 + a_2\xi_1 + a_3\xi_2 + a_4\xi_1\xi_2$$

with respect to ξ_1 and ξ_2 is used for the basis function. Four undetermined multipliers a_1, \dots, a_4 are determined by the boundary conditions of four nodes of the normal element:

$$(\xi_{(1)} \ \xi_{(2)} \ \xi_{(3)} \ \xi_{(4)}) = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}. \quad (6.4.24)$$

In reality, using the conditions $\hat{\varphi}_{(\alpha)}(\xi_{(\beta)}) = \delta_{\alpha\beta}$ with respect to $\alpha, \beta \in \{1, 2, 3, 4\}$,

$$(\hat{\varphi}_{(1)} \ \hat{\varphi}_{(2)} \ \hat{\varphi}_{(3)} \ \hat{\varphi}_{(4)}) = (\lambda_{11}\lambda_{21} \ \lambda_{12}\lambda_{21} \ \lambda_{12}\lambda_{22} \ \lambda_{11}\lambda_{22})$$

can be obtained as the basis functions of a normal element. Figure 6.27 shows these basis functions. Using these, the approximate function of the normal element can be constructed as Eq. (6.4.9).

Two methods are known for forming higher-order rectangular finite elements. One is a method such as the one below using bi- m -th order polynomials.

Definition 6.4.3 (Lagrange Family Rectangular Finite Element) Let $\Xi = (0, 1)^2$ be a normal domain. For $m \in \mathbb{N}$, nodes $\xi_{(\alpha)}$ with respect to $\alpha \in \mathcal{N}_i$ are placed as in Fig. 6.28. A basis function $\hat{\varphi}_{(\alpha)}$ is constructed using bi- m -th order polynomials with respect to ξ_1 and ξ_2 and undetermined multipliers are determined so that $\hat{\varphi}_{(\alpha)}(\xi_{(\beta)}) = \delta_{\alpha\beta}$ is satisfied with respect to $\alpha, \beta \in \mathcal{N}_i$. A finite element using the basis functions constructed in this way is called a Lagrange family rectangular finite element. \square

Another method for forming a higher order is to use just the nodes on the finite element boundary. This method is different from the Lagrange family which was obtained by deduction using the bi- m -th order polynomials and, from the fact that the method was found accidentally, it is called the serendipity group.

Fig. 6.27 Basis functions and approximate function used in a first-order rectangular finite element

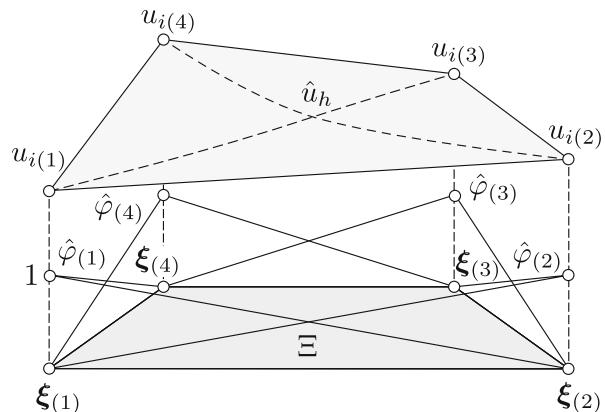


Fig. 6.28 Nodes of a Lagrange family rectangular finite element

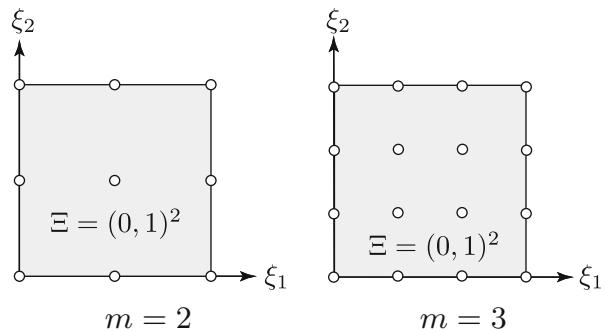
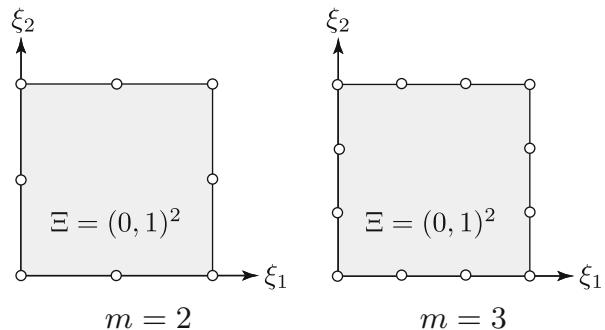


Fig. 6.29 Nodes of a serendipity group rectangular finite element



Definition 6.4.4 (Serendipity Group Rectangular Finite Element) Let $\Xi = (0, 1)^2$ be a normal domain. For $m \in \mathbb{N}$, place the nodes $\xi_{(\alpha)}$ with respect to $\alpha \in \mathcal{N}_i$ as shown in Fig. 6.29. Construct the basis functions $\hat{\varphi}_{(\alpha)}$ using polynomials of ξ_1 and ξ_2 and determine the undetermined multipliers so that $\hat{\varphi}_{(\alpha)}(\xi_{(\beta)}) = \delta_{\alpha\beta}$ is satisfied with respect to $\alpha, \beta \in \mathcal{N}_i$. A finite element using these basis functions created in this way is called a serendipity group rectangular finite element. These polynomials are constructed by terms which are complete second-order polynomials with respect to

ξ_1 and ξ_2 with $\xi_1^2 \xi_2$ and $\xi_1 \xi_2^2$ added on when $m = 2$. Moreover, when $m = 3$, these are constructed by $\xi_1^3 \xi_2$ and $\xi_1 \xi_2^3$ being added to complete third-order polynomials with respect to ξ_1 and ξ_2 . \square

As shown above, if the approximate function $\hat{u}_h(\xi)$ is given on the standard coordinate $\xi \in \Xi$, the affine mapping from normal coordinates Ξ to rectangular finite element Ω_i can be used to obtain an element coefficient matrix or known term vector of Ω_i via Eqs. (6.4.15) and (6.4.17). In this case, when the rectangular element is as shown in Fig. 6.26 and $\mathbf{x}_{i(\alpha)} = (x_{i(\alpha)1}, x_{i(\alpha)2})^\top$ with respect to $\alpha \in \{1, 2, 3, 4\}$ are taken to be the coordinate values of local nodes of the finite element $i \in \mathcal{E}$, the affine mapping is given by

$$\begin{aligned} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &= \begin{pmatrix} x_{i(2)1} - x_{i(1)1} & 0 \\ 0 & x_{i(4)2} - x_{i(1)2} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} x_{i(1)1} \\ x_{i(1)2} \end{pmatrix} \\ &= \mathbf{F}_i \xi + \mathbf{x}_{i(1)}. \end{aligned} \quad (6.4.25)$$

6.4.4 Tetrahedral Finite Elements

A three-dimensional finite element can also be constructed in a similar way to the two-dimensional case. First, let us consider a tetrahedron finite element such as the one in Fig. 6.30. Area coordinates were used in triangular finite elements. In a tetrahedral finite element, volume coordinates $(\lambda_1, \dots, \lambda_4)$ such as that in Fig. 6.31a and a normal domain $\Xi = \{\xi \in (0, 1)^3 \mid \xi_1 + \xi_2 + \xi_3 < 1\}$ such as the one in Fig. 6.31b are used. In first-order tetrahedral finite elements, $\hat{\varphi}_{(1)} = \lambda_1, \dots, \hat{\varphi}_{(1)} = \lambda_4$ with respect to the nodes $\xi_{(1)}, \dots, \xi_{(4)}$ such as the one in Fig. 6.31b are chosen to be basis functions. The m -th order tetrahedron finite element has basis functions using complete m -th order polynomials in a similar way to triangular finite elements.

Fig. 6.30 First-order tetrahedral finite element

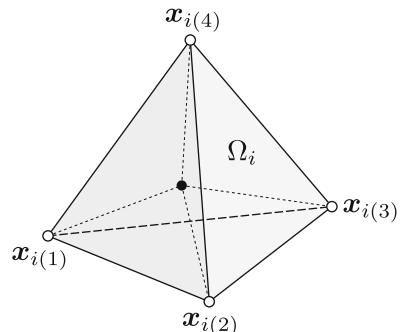


Fig. 6.31 Volume coordinates and normal coordinates for a tetrahedral finite element. (a) Volume coordinates $(\lambda_1, \dots, \lambda_4)$. (b) Normal coordinates (ξ_1, ξ_2, ξ_3)

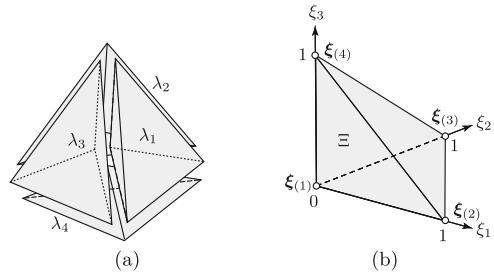
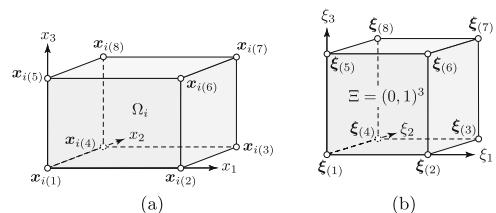


Fig. 6.32 First-order hexahedral finite element and normal coordinates. (a) Hexahedral first-order finite element. (b) Normal coordinates (ξ_1, ξ_2, ξ_3)



6.4.5 Hexahedral Finite Elements

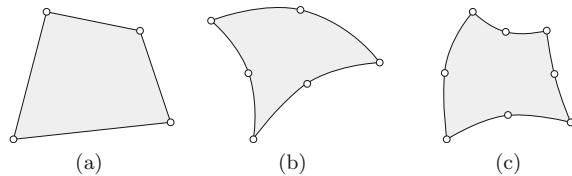
A hexahedral finite element is also constructed by expanding the rectangular finite element into a three-dimensional space. Figure 6.32 shows nodes of a first-order hexahedral finite element. Here, it is also possible, in a similar way to Eq. (6.4.20) to Eq. (6.4.23), to define the length coordinates $\lambda_{11}, \dots, \lambda_{33}$ and normal coordinates $\xi = (\xi_1, \xi_2, \xi_3)^\top = (\lambda_{12}, \lambda_{22}, \lambda_{32})^\top \in \Xi$ with respect to the node coordinates $\mathbf{x}_{i(1)}, \dots, \mathbf{x}_{i(8)} \in \mathbb{R}^3$. With respect to a normal element, the Lagrange family m -th order hexahedral finite element has nodes uniformly placed on the normal domain and basis functions $\hat{\varphi}_{(1)}, \dots, \hat{\varphi}_{((m+1)^3)}$ are constructed using tri- m -th order polynomials with respect to normal coordinates. In serendipity group m -th order hexahedral finite elements, nodes are placed uniformly on the finite element boundary and basis functions constructed using m -th order polynomials.

6.5 Isoparametric Finite Elements

The domain of a finite element seen in Sect. 6.4 is assumed to be triangular or rectangular in two dimensions and tetrahedral or hexahedral in three dimensions. Here, let us think about finite elements in the shape of a quadrangle, triangle or quadrangle formed of second-order curves such as shown in Fig. 6.33 and those extended to the three dimensions.

As seen in Sect. 6.4.2, when the basis functions of a triangular finite element are given by area coordinates and the approximate functions constructed from these are substituted into the weak form, the domain integrals are calculated using

Fig. 6.33 Examples of isoparametric finite elements.
 (a) Quadrangle. (b) 2nd-order curve triangular. (c) 2nd-order curve quadrangle



Theorem 6.3.1. In the case of rectangular finite elements, they can be calculated using Gaussian quadrature, as shown in Sect. 6.5.2. These formulae of integration are also effective when the order number of the integrand is increased by making the basis functions a higher order. However, if the integral domain is as shown in Fig. 6.33, such integral formulae can no longer be used.

Hence, if a function (mapping) is used for changing the finite element domain Ω_i into a triangular or rectangular normal domain Ξ if two-dimensional and into a tetrahedral or hexahedral normal coordinates Ξ if three-dimensional, such a change makes integration possible using Theorem 6.3.1 or Gaussian quadrature on Ξ . When the approximate function \hat{x}_h with respect to the mapping $\hat{x} : \Xi \rightarrow \Omega_i$ is constructed of the same basis functions as the approximate function of u , the finite element is called an isoparametric finite element. In other words, it is defined in the following way.

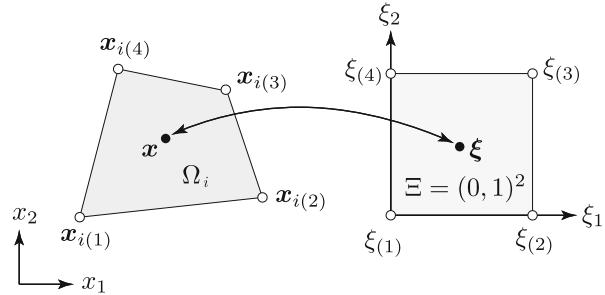
Definition 6.5.1 (Isoparametric Finite Element) Let the normal domain with respect to the domain $\Omega_i \subset \mathbb{R}^d$ of finite element $i \in \mathcal{E}$ be $\Xi \subset \mathbb{R}^d$. With respect to local node number $i \in \mathcal{N}_i = \{1, \dots, |\mathcal{N}_i|\}$, the basis function of the normal element is $\hat{\phi} = (\hat{\phi}_{(1)}, \dots, \hat{\phi}_{(|\mathcal{N}_i|)})$. In this case, the finite element when the approximate functions of u and v and coordinate values on Ω_i are constructed using

$$\begin{aligned}\hat{u}_h(\xi) &= \hat{\phi}(\xi) \cdot \bar{u}_i, \\ \hat{v}_h(\xi) &= \hat{\phi}(\xi) \cdot \bar{v}_i, \\ \hat{x}_{h1}(\xi) &= \hat{\phi}(\xi) \cdot \bar{x}_{i1}, \\ &\vdots \\ \hat{x}_{hd}(\xi) &= \hat{\phi}(\xi) \cdot \bar{x}_{id}\end{aligned}$$

with respect to $\xi \in \Xi$ is called an isoparametric finite element. Here, \bar{u}_i and $\bar{v}_i \in \mathbb{R}^{|\mathcal{N}_i|}$ are taken to be local node value vectors of u and v , and $\bar{x}_{i1}, \dots, \bar{x}_{id} \in \mathbb{R}^{|\mathcal{N}_i|}$ are taken to be local node coordinate value vectors on Ω_i . \square

In an isoparametric finite element, all functions appearing in the weak form with respect to a finite element are given by the normal coordinates $\xi \in \Xi$ as parameters. As a result, the integral domain can be changed to a normal domain and the formula of integration can be used. However, in contrast, the calculation of the partial differential of u with respect to $x \in \Omega_i$ and the Jacobian may be difficult. Let us look at this in the next section.

Fig. 6.34 Coordinate transformation of four-node isoparametric finite element



6.5.1 Two-Dimensional Four-Node Isoparametric Finite Elements

As an example of an isoparametric finite element, let us think about a four-node isoparametric finite element such as the one in Fig. 6.34. With respect to $i \in \mathcal{E}$, assume Ω_i to be a quadrangle domain and $\Xi = (0, 1)^2$ a normal domain. Here, with respect to $\xi \in \Xi$,

$$\hat{u}_h(\xi) = (\hat{\varphi}_{(1)}(\xi) \ \hat{\varphi}_{(2)}(\xi) \ \hat{\varphi}_{(3)}(\xi) \ \hat{\varphi}_{(4)}(\xi)) \begin{pmatrix} u_{i(1)} \\ u_{i(2)} \\ u_{i(3)} \\ u_{i(4)} \end{pmatrix} = \hat{\varphi}(\xi) \cdot \bar{u}_i,$$

$$\hat{v}_h(\xi) = \hat{\varphi}(\xi) \cdot \bar{v}_i,$$

$$\hat{x}_{h1}(\xi) = \hat{\varphi}(\xi) \cdot \bar{x}_{i1},$$

$$\hat{x}_{h2}(\xi) = \hat{\varphi}(\xi) \cdot \bar{x}_{i2}$$

is defined, where

$$\hat{\varphi} = \begin{pmatrix} \hat{\varphi}_{(1)} \\ \hat{\varphi}_{(2)} \\ \hat{\varphi}_{(3)} \\ \hat{\varphi}_{(4)} \end{pmatrix} = \begin{pmatrix} (1 - \xi_1)(1 - \xi_2) \\ \xi_1(1 - \xi_2) \\ \xi_1\xi_2 \\ (1 - \xi_1)\xi_2 \end{pmatrix}.$$

Here, let us think about the calculation of partial differentials with respect to x_1 and x_2 of $\hat{\varphi}$. However, $\hat{\varphi}$ is a function of ξ . Here, if the chain rule of differentiation is used, with respect to $\alpha \in \{1, 2, 3, 4\}$,

$$\begin{aligned} \nabla_{\xi} \hat{\varphi}_{(\alpha)}(\xi) &= \begin{pmatrix} \partial \hat{\varphi}_{(\alpha)} / \partial \xi_1 \\ \partial \hat{\varphi}_{(\alpha)} / \partial \xi_2 \end{pmatrix} = \begin{pmatrix} \partial \hat{x}_1 / \partial \xi_1 & \partial \hat{x}_2 / \partial \xi_1 \\ \partial \hat{x}_1 / \partial \xi_2 & \partial \hat{x}_2 / \partial \xi_2 \end{pmatrix} \begin{pmatrix} \partial \hat{\varphi}_{(\alpha)} / \partial x_1 \\ \partial \hat{\varphi}_{(\alpha)} / \partial x_2 \end{pmatrix} \\ &= (\nabla_{\xi} \hat{x}^{\top}) \nabla_x \hat{\varphi}_{(\alpha)}(\xi) \end{aligned}$$

is established. Then,

$$\begin{aligned}
 \nabla_x \hat{\varphi}_{(\alpha)}(\xi) &= \begin{pmatrix} \partial \hat{\varphi}_{(\alpha)} / \partial x_1 \\ \partial \hat{\varphi}_{(\alpha)} / \partial x_2 \end{pmatrix} \\
 &= \frac{1}{\omega_i(\xi)} \begin{pmatrix} \partial \hat{x}_2 / \partial \xi_2 & -\partial \hat{x}_2 / \partial \xi_1 \\ -\partial \hat{x}_1 / \partial \xi_2 & \partial \hat{x}_1 / \partial \xi_1 \end{pmatrix} \begin{pmatrix} \partial \hat{\varphi}_{(\alpha)} / \partial \xi_1 \\ \partial \hat{\varphi}_{(\alpha)} / \partial \xi_2 \end{pmatrix} \\
 &= \left(\nabla_\xi \hat{\mathbf{x}}^\top \right)^{-1} \nabla_\xi \hat{\varphi}_{(\alpha)}(\xi)
 \end{aligned} \tag{6.5.1}$$

is established. Here, $\left(\nabla_\xi \hat{\mathbf{x}}^\top \right)^\top$ and

$$\omega_i(\xi) = \det \left(\nabla_\xi \hat{\mathbf{x}}^\top \right) \tag{6.5.2}$$

are the Jacobi matrix and the Jacobian, respectively, of the mapping $\hat{\mathbf{x}} : \Xi \rightarrow \Omega_i$.

Using these results, the element coefficient matrix $(a_i(\varphi_{i(\alpha)}, \varphi_{i(\beta)}))_{\alpha, \beta} \in \mathbb{R}^{4 \times 4}$ can be changed in the following way:

$$\begin{aligned}
 a_i(\varphi_{i(\alpha)}, \varphi_{i(\beta)}) &= \int_{\Omega_i} \left(\frac{\partial \varphi_{i(\alpha)}}{\partial x_1} \frac{\partial \varphi_{i(\beta)}}{\partial x_1} + \frac{\partial \varphi_{i(\alpha)}}{\partial x_2} \frac{\partial \varphi_{i(\beta)}}{\partial x_2} \right) dx \\
 &= \int_{\Omega_i} \nabla_x \varphi_{i(\alpha)}(\mathbf{x}) \cdot \nabla_x \varphi_{i(\beta)}(\mathbf{x}) dx \\
 &= \int_{(0,1)^2} \nabla_\xi \hat{\varphi}_{(\alpha)}(\xi) \cdot \nabla_\xi \hat{\varphi}_{(\beta)}(\xi) \omega_i(\xi) d\xi.
 \end{aligned} \tag{6.5.3}$$

In the integrand on the right-hand side of Eq. (6.5.3), $\nabla_x \hat{\varphi}_{(\alpha)}(\xi)$ and $\omega_i(\xi)$ can be calculated by Eqs. (6.5.1) and (6.5.2), respectively. The integral can be calculated with the Gaussian quadrature shown next.

6.5.2 Gaussian Quadrature

Let us show the formula of Gaussian quadrature specifically. When $f_n : (-1, 1) \rightarrow \mathbb{R}$ is n -th order function with respect to $n \in \mathbb{N}$, using the fact that

$$\begin{aligned}
 \int_{-1}^1 f_1(y) dy &= 2f_1(0), \\
 \int_{-1}^1 f_3(y) dy &= f_3\left(-\frac{1}{\sqrt{3}}\right) + f_3\left(\frac{1}{\sqrt{3}}\right),
 \end{aligned}$$

$$\int_{-1}^1 f_5(y) dy = \frac{5}{9} f_5 \left(-\sqrt{\frac{3}{5}} \right) + \frac{8}{9} f_5(0) + \frac{5}{9} f_5 \left(\sqrt{\frac{3}{5}} \right),$$

$$\vdots$$

holds for $n \in \{1, 3, 5, \dots\}$, the method for calculating the integration on the left-hand side using the right-hand side is called the Gaussian quadrature. Here, when the term on the right-hand side is written with respect to $i \in \{1, 2, \dots, (n+1)/2\}$ as $w_i f_n(\eta_i)$, η_i is called the Gaussian node. Figure 6.35 shows the relationship between f_n and Gaussian nodes. Let us see the basis on which these formulae hold.

First, in order to use the Gaussian quadrature theorem to be shown later, let us define Legendre polynomials. Here, n and m are non-negative integers.

Definition 6.5.2 (Legendre Polynomials) When the function $l_n : (-1, 1) \rightarrow \mathbb{R}$ satisfies the Legendre differential equation:

$$\frac{d}{dx} \left\{ (1-x^2) \frac{d}{dx} l_n \right\} + n(n+1) l_n = 0, \quad (6.5.4)$$

l_n is called a Legendre polynomial. \square

Using Rodrigues' formula, the Legendre polynomial can be written as

$$l_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} \left\{ (x^2 - 1)^n \right\}. \quad (6.5.5)$$

From this, it can be obtained specifically as

$$l_0(x) = 1, \quad l_1(x) = x, \quad l_2(x) = x^2 - \frac{1}{3}, \quad l_3(x) = x^3 - \frac{3}{5}x, \quad \dots.$$

Figure 6.36 shows l_0 to l_5 . Therefore, it becomes apparent that l_n is an n -th order polynomial. Furthermore, if the function $f : (-1, 1) \rightarrow \mathbb{R}$ is a polynomial of less than n -th order then

$$\int_{-1}^1 f(x) l_n(x) dx = 0$$

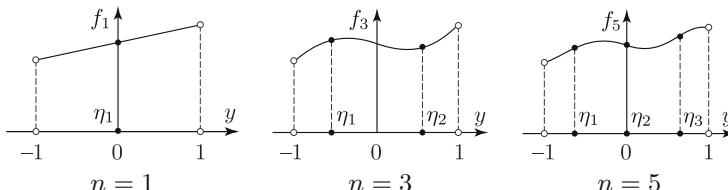
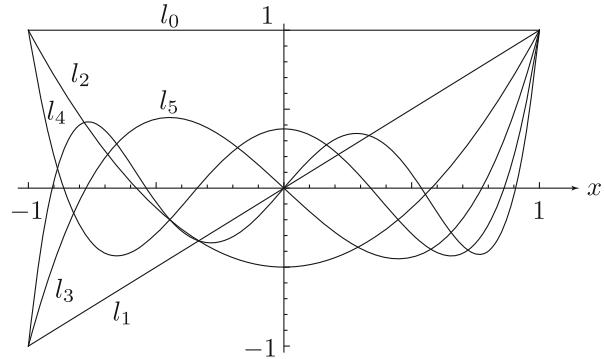


Fig. 6.35 Gaussian quadrature of a one-dimensional function

Fig. 6.36 Legendre polynomial l_n



holds. This is because

$$\begin{aligned} \int_{-1}^1 f(x) l_n(x) dx &= \left[f \left\{ -\frac{1}{n(n+1)} (1-x^2) \frac{d l_n}{dx} \right\} \right]_{-1}^1 \\ &\quad - \frac{1}{2^n n!} \int_{-1}^1 \frac{d f}{dx} \frac{d^{n-1}}{dx^{n-1}} \left\{ (x^2 - 1)^n \right\} dx \\ &= \frac{(-1)^n}{2^n n!} \int_{-1}^1 \frac{d^n f}{dx^n} (x^2 - 1)^n dx = 0 \end{aligned}$$

holds from Eqs. (6.5.4) and (6.5.5). Using these properties, the orthogonality of $\{l_n\}_n$:

$$\int_{-1}^1 l_n(x) l_m(x) dx = \frac{2}{2n+1} \delta_{nm}$$

is established.

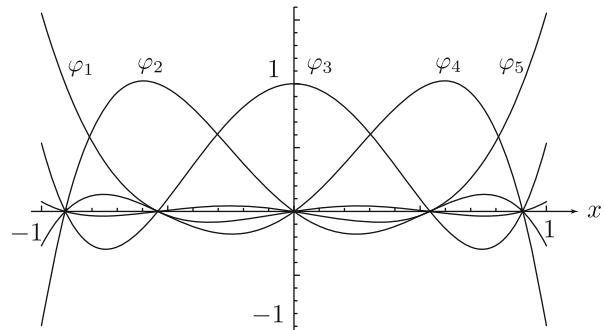
Moreover, in general, with respect to $x_0 < x_1 < \dots < x_n$,

$$\begin{aligned} \phi_i(x) &= \prod_{j \in \{1, \dots, n\}, j \neq i} \frac{x - x_j}{x_i - x_j} \\ &= \frac{(x - x_0)(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)} \end{aligned}$$

satisfies $\phi_i(x_j) = \delta_{ij}$, and with respect to some $f : \mathbb{R} \rightarrow \mathbb{R}$, if we set

$$\hat{f}(x) = \sum_{i \in \{1, \dots, n\}} \phi_i(x) f(x_i),$$

Fig. 6.37 Functions $\varphi_i(x)$ with the roots of a fifth-order Legendre polynomial as nodes



then the equation $\hat{f}(x_i) = f(x_i)$ is satisfied. Here, $\phi_i(x)$ is called the Lagrange basis polynomials and $\hat{f}(x)$ is Lagrange interpolation. At this time, we write the Lagrange basis polynomials with respect to the roots η_1, \dots, η_n of Legendre polynomial l_n as

$$\varphi_i(x) = \prod_{j \in \{1, \dots, n\}, j \neq i} \frac{x - \eta_j}{\eta_i - \eta_j}. \quad (6.5.6)$$

Figure 6.37 shows φ_1 to φ_5 . If we consider Lagrange interpolation using $\varphi_i(x)$, a formula of Gaussian quadrature can be obtained as follows.

Theorem 6.5.3 (Gaussian Quadrature) *Let η_1, \dots, η_n be the roots of an n -th order Legendre polynomial l_n . Let $f : (-1, 1) \rightarrow \mathbb{R}$ be a polynomial of less than $2n$ -th order. In this case,*

$$\int_{-1}^1 f(x) dx = \sum_{i \in \{1, \dots, n\}} w_i f(\eta_i)$$

holds. Here,

$$w_i = \int_{-1}^1 \varphi_i(x) dx$$

with respect to $\varphi_i(x)$ of Eq. (6.5.6). □

Proof Let us suppose that $f(x)$ is a polynomial of $(n-1)$ -th order. In this case,

$$f(x) = \sum_{i \in \{1, \dots, n\}} \varphi_i(x) f(\eta_i)$$

holds. This is because although the difference between both sides includes an n -th order differential but f 's n -th order differential is 0. Hence, the following holds:

$$\int_{-1}^1 f(x) dx = \int_{-1}^1 \left(\sum_{i \in \{1, \dots, n\}} \varphi_i(x) f(\eta_i) \right) dx = \sum_{i \in \{1, \dots, n\}} w_i f(\eta_i).$$

Next, let us suppose f is a polynomial of order greater than n but less than $2n$. In this case we can write

$$f(x) = l_n(x) g(x) + r(x).$$

However, $g(x)$ and $r(x)$ are polynomials of less than n -th order. Here, the qualities of the Legendre polynomial give

$$\int_{-1}^1 l_n(x) g(x) dx = 0.$$

Moreover, from $l_n(\eta_i) = 0$,

$$f(\eta_i) = r(\eta_i)$$

holds. Furthermore, since $r(x)$ is a polynomial of less than n -th order,

$$\int_{-1}^1 r(x) dx = \sum_{i \in \{1, \dots, n\}} w_i r(\eta_i)$$

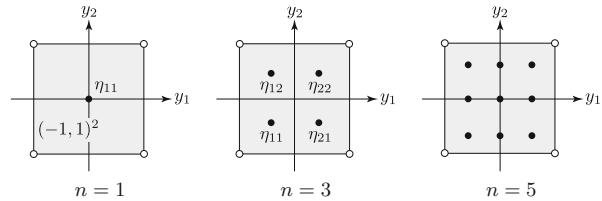
holds. Therefore we get

$$\int_{-1}^1 f(x) dx = \int_{-1}^1 r(x) dx = \sum_{i \in \{1, \dots, n\}} w_i r(\eta_i) = \sum_{i \in \{1, \dots, n\}} w_i f(\eta_i). \quad \square$$

In Theorem 6.5.3, when the integral domain is changed to $(0, 1)$, the integral equation can be changed by a change in variables to

$$\int_0^1 f_{2n-1}(y) dy = \frac{1}{2} \sum_{i \in \{1, \dots, n\}} w_i f\left(\frac{\eta_i - 1}{2}\right).$$

Fig. 6.38 Gaussian quadrature for functions defined on a two-dimensional domain



Furthermore, with respect to a bi- n -th order function on two-dimensional domain $(-1, 1)^2$,

$$\int_{(-1,1)^2} f_{2n-1}(\xi) \, d\xi = \sum_{(i,j) \in \{1, \dots, n\}^2} w_i w_j f(\eta_{ij}),$$

$$\int_{(0,1)^2} f_{2n-1}(\xi) \, d\xi = \frac{1}{4} \sum_{(i,j) \in \{1, \dots, n\}^2} w_i w_j f\left(\left(\eta_{ij} - \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) / 2\right) \quad (6.5.7)$$

holds with respect to Gaussian nodes such as in Fig. 6.38. These formulae are used in numerical integration of rectangular isoparametric finite elements. For example, with respect to Eq. (6.5.3), Eq. (6.5.7) is used. Here, the value of n is not chosen exactly due to the fact that the inverse calculation of a matrix with polynomials as elements is included. Hence a value as small as possible such that no practical issues arise can be looked for with a numerical experiment. Moreover, with respect to linear elastic problems, it is known that the use of selected reduced integration suppresses the generation of hourglass modes (deformation such that strain becomes 0). We refer the readers to literature focusing on these topics.

6.6 Error Estimation

In Theorem 6.1.13, it was seen that the approximate solution from the Galerkin method was the best element in the set U_h of approximate functions. Hence, the error in the approximate solution from the Galerkin method depends on how close an element of U_h gets to an element in the function space containing the exact solution (approximation ability). Here, the results of Theorem 6.1.13 will be used to think about the error evaluation of an approximate solution (finite element solution) obtained from the finite element method. Results shown here will be used for error evaluation in numerical solutions for shape optimization problems in Chaps. 8 and 9. Here, based on what we have seen so far in this chapter, let us give an abstraction of the finite element method to some extent to look at basic theorems.

6.6.1 Finite Element Division Sequence

Let us define the finite element division. Let $\Omega \subset \mathbb{R}^d$ be a $d \in \{1, 2, 3\}$ -dimensional bounded domain of a polygon in two dimensions and a polyhedron in three dimensions in order to be able to ignore the error due to domain division, and call it polyhedron generally. With respect to Ω , $\mathcal{T} = \{\Omega_i\}_{i \in \mathcal{E}}$ is called a finite element division. Here, \mathcal{E} is the finite set of element numbers. Moreover, Ω_i are convex polyhedrons such that $\bar{\Omega} = \bigcup_{i \in \mathcal{E}} \bar{\Omega}_i$ and

$$\Omega_i \cap \Omega_j = \emptyset, \quad \bar{\Omega}_i \cap \bar{\Omega}_j \subset \mathbb{R}^{d-1}$$

for $i \neq j$ with respect to an arbitrary $i, j \in \mathcal{E}$ is satisfied. With respect to each Ω_i , we call

$$\text{diam } \Omega_i = \sup_{\mathbf{x}, \mathbf{y} \in \Omega_i} \|\mathbf{x} - \mathbf{y}\|_{\mathbb{R}^d}$$

the diameter of Ω_i and write it as h_i . Moreover, write the diameter of the inscribed sphere in Ω_i as $\text{inscr } \Omega_i$. When some positive real number σ exists and

$$\frac{\text{inscr } \Omega_i}{\text{diam } \Omega_i} \geq \sigma \quad (6.6.1)$$

with respect to all $i \in \mathcal{E}$ holds, \mathcal{T} is said to be a regular finite element division. Furthermore,

$$h(\mathcal{T}) = \max_{i \in \mathcal{E}} h_i \quad (6.6.2)$$

is referred to as the maximum diameter of \mathcal{T} and a finite element division sequence such that it becomes $h(\mathcal{T}) \rightarrow 0$ will be written as $\{\mathcal{T}_h\}_{h \rightarrow 0}$.

6.6.2 Affine-Equivalent Finite Element Division Sequence

Choose some finite element division \mathcal{T}_h from a regular finite element division sequence $\{\mathcal{T}_h\}_{h \rightarrow 0}$ and let the set of node numbers in this case be \mathcal{N} and \mathbf{x}_j a node with respect to $j \in \mathcal{N}$. Here, the basis functions $\phi_j : \Omega \rightarrow \mathbb{R}$ of the finite element method when viewed as a Galerkin method are assumed to satisfy the following conditions:

(1) With respect to an arbitrary $j, l \in \mathcal{N}$,

$$\phi_j(\mathbf{x}_l) = \delta_{jl}$$

holds.

(2) ϕ_j has support on a domain of finite elements with node $j \in \mathcal{N}$.

The set of these basis functions ϕ_j , as seen in Sects. 6.2 to 6.5, can be rewritten as a set of basis functions $\varphi_{i(\alpha)}$ ($\alpha \in \mathcal{N}_i$) defined on the domain Ω_i of finite elements $i \in \mathcal{E}$. Furthermore, we set the normal domain to be Ξ , basis functions on Ξ given by $\hat{\varphi}_{(\alpha)} : \Xi \rightarrow \mathbb{R}$, the mapping of normal domain to finite element domain is written as $f_i : \Xi \rightarrow \Omega_i$, and

$$\varphi_{i(\alpha)}(f_i(\xi)) = \hat{\varphi}(\xi_{(\alpha)}).$$

In this section, for simplicity, all normal elements corresponding to all finite elements are taken to be common as

$$\Xi = (0, 1)^d \quad \text{or} \quad \Xi = \left\{ \xi \in (0, 1)^d \mid \xi_1 + \dots + \xi_d < 1 \right\}. \quad (6.6.3)$$

Moreover, f_i is assumed to be given by a linear form such as

$$f_i(\xi) = \mathbf{F}_i \xi + \mathbf{b}_i \quad (6.6.4)$$

using $\mathbf{F}_i \in \mathbb{R}^{d \times d}$ and $\mathbf{b}_i \in \mathbb{R}^d$. In the case of triangular or rectangular finite elements, they are specifically shown in Eqs. (6.4.16) and (6.4.25). A mapping such as this, combining a linear mapping and a translation with a fixed element, is called an affine mapping. Here, the finite element division when f_i is an affine mapping is called an affine-equivalent finite element division.

The finite element division seen from Sects. 6.2 to 6.4 had assumed an affine-equivalent finite element division sequence. With isoparametric finite elements, this relationship does not generally hold, but even if $f_i : \Xi \rightarrow \Omega_i$ is a mapping combining a non-linear mapping and a translation with a fixed element, if the Jacobi matrix $f_{i\xi^\top} = (\nabla_\xi f_i^\top)^\top$ and Jacobian $\omega_i(\xi) = \det f_{i\xi^\top}$ of f_i shown later have upper limits and lower limits given by positive real numbers, then a similar argument holds.

When using a regular affine-equivalent finite element division sequence defined in this way, in a Poisson problem for example, there is the need to calculate

$$\begin{aligned} a_i(\varphi_{i(\alpha)}, \varphi_{i(\beta)}) &= \int_{\Omega_i} \nabla_x \varphi_{i(\alpha)} \cdot \nabla_x \varphi_{i(\beta)} \, dx \\ &= \int_{\Xi} \left\{ (\nabla_\xi f_i^\top)^{-1} \nabla_\xi \hat{\varphi}_{(\alpha)} \right\} \cdot \left\{ (\nabla_\xi f_i^\top)^{-1} \nabla_\xi \hat{\varphi}_{(\beta)} \right\} \omega_i \, d\xi \\ &= \int_{\Xi} \left\{ (\mathbf{F}_i^\top)^{-1} \nabla_\xi \hat{\varphi}_{(\alpha)} \right\} \cdot \left\{ (\mathbf{F}_i^\top)^{-1} \nabla_\xi \hat{\varphi}_{(\beta)} \right\} \omega_i \, d\xi \end{aligned} \quad (6.6.5)$$

with respect to an arbitrary $\alpha, \beta \in \mathcal{N}_i$, where \mathbf{F}_i is a matrix used in Eq. (6.6.4) and $\omega_i = \det \mathbf{F}_i$. The following results can be obtained with respect to \mathbf{F}_i and ω_i of Eq. (6.6.5).

Lemma 6.6.1 (Jacobian of an Affine Mapping) *Let $\mathcal{T}_h = \{\Omega_i\}_{i \in \mathcal{E}}$ be a regular affine-equivalent finite element division sequence with normal domain Ξ as Eq. (6.6.3) and affine mapping \mathbf{f}_i as Eq. (6.6.4). Let $\text{diam } \Omega_i = h_i$. In this case,*

$$\omega_i = \det \mathbf{F}_i \leq h_i^d, \quad \omega_i^{-1} = \det(\mathbf{F}_i)^{-1} \leq c_1 h_i^{-d} \quad (6.6.6)$$

holds. Moreover, with respect to $\mathbf{F}_i = (a_{ijk})_{jk} \in \mathbb{R}^{|\mathcal{N}_i| \times |\mathcal{N}_i|}$ and $\mathbf{F}_i^{-1} = (a_{ijk}^{-1})_{jk} \in \mathbb{R}^{|\mathcal{N}_i| \times |\mathcal{N}_i|}$,

$$|a_{ijk}| \leq h_i, \quad |a_{ijk}^{-1}| \leq c_2 h_i^{-1} \quad (6.6.7)$$

is established. Here, c_1 and c_2 is positive constants depending on σ in Eq. (6.6.1) and d . \square

Proof If it is a finite element of an affine-equivalent finite element division sequence, ω_i is a constant. When Ξ is Eq. (6.6.3), $0 < |\Xi| \leq 1$. \mathcal{T}_h is regular, hence

$$\frac{1}{c_1} h_i^d \leq \omega_i = \frac{\int_{\Xi} \omega_i \, d\xi}{\int_{\Xi} d\xi} = \frac{|\Omega_i|}{|\Xi|} \leq h_i^d.$$

In other words, Eq. (6.6.6) holds. Moreover, if $\mathbf{f}_i(\xi)$ of Eq. (6.6.4) is written as $(f_{ij}(\xi))_j$,

$$|a_{ijk}| = \left| \frac{\partial f_{ij}}{\partial \xi_k} \right| \leq h_i$$

is obtained from $\text{diam } \Omega_i = h_i$. Furthermore, from

$$\omega_i^{-1} = \det \left(\frac{\partial f_{ij}^{-1}}{\partial x_k} \right)_{jk} \leq \frac{c_1}{h_i^d},$$

the following is obtained:

$$|a_{ijk}^{-1}| = \left| \frac{\partial f_{ij}^{-1}}{\partial x_k} \right| \leq \frac{c_2}{h_i}.$$

In other words, Eq. (6.6.7) is established. \square

6.6.3 Interpolation Error Estimation

In Sect. 6.6.2, attention was given to the relationship between the normal element and the finite element. Here, let us focus on the approximation ability of the approximate function. In Theorem 6.1.13, the finite element solution error $\|u - u_h\|_U$ measured with a Hilbert space containing the exact solution was seen to be limited to $\inf_{v_h \in U_h} \|u - v_h\|_U$ at the homogeneous basic boundary condition ($u_D = 0$). Here, in order to evaluate its approximation ability, we shall think about creating an approximate function for which it is easy to evaluate error, but the error may be greater than for the finite element solution. If an error of such an approximate function is evaluated, from the fact that the error of the finite element solution is smaller (Theorem 6.1.13), the error of the finite element solution should be evaluated using that error. This is shown in Sect. 6.6.4.

In preparation, in this section, let us think about approximate functions which can be easily evaluated for error. Such an approximate function is assumed to be a function which is an element of U_h and agrees with the exact solution at nodes. Such an approximate function is called an interpolation function. Figure 6.39a shows the relationship between the exact solution u , the finite element solution u_h and the interpolation function πu when basis functions are given by linear functions with respect to a one-dimensional problem. Figure 6.39b shows the functions defined on the normal elements. Here, π and $\hat{\pi}$ are called interpolation operators and are defined as below. Write the function space of exact solutions on Ω_i as $U(\Omega_i)$. Moreover, the function space (linear space) of an interpolation function set by the basis functions $\varphi_{i(\alpha)}$ ($\alpha \in \mathcal{N}_i$) is written as $W(\Omega_i) = \text{span}(\varphi_{i(\alpha)})_{\alpha \in \mathcal{N}_i}$ (Definition 4.2.6). On the other hand, the function spaces of the exact solution and interpolation functions defined on Ξ are respectively written as $U(\Xi)$ and $W(\Xi) = \text{span}(\hat{\varphi}_{(\alpha)})_{\alpha \in \mathcal{N}_i}$. Here, the operators $\pi : U(\Omega_i) \rightarrow W(\Omega_i)$ and

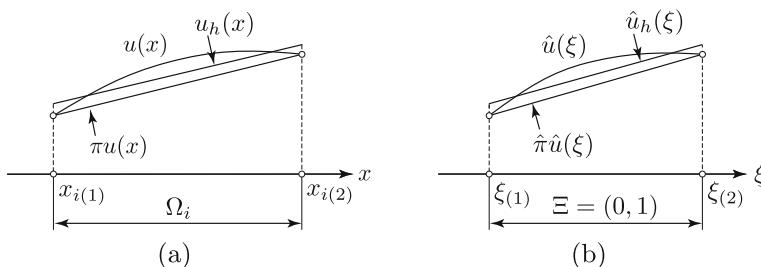


Fig. 6.39 Interpolation function and finite element solution. (a) Interpolation function πu on Ω_i . (b) Interpolation function $\hat{\pi} \hat{u}$ on Ξ

$\hat{\pi} : U(\Xi) \rightarrow W(\Xi)$ are defined by

$$\pi u(\mathbf{x}) = \sum_{\alpha \in \mathcal{N}_i} u(\mathbf{x}_{i(\alpha)}) \varphi_{i(\alpha)}(\mathbf{x}), \quad (6.6.8)$$

$$\hat{\pi} \hat{u}(\xi) = \sum_{\alpha \in \mathcal{N}_i} \hat{u}(\xi_{(\alpha)}) \hat{\varphi}_{(\alpha)}(\xi). \quad (6.6.9)$$

Such an error $\|u - \pi u\|_U$, that is $\|u - \pi u\|_{H^1(\Omega; \mathbb{R})}$, in an interpolation function is called an interpolation error. In order to evaluate this, first let us define the set of all k -th order polynomials and look at results relating to their general approximation abilities. For $k \in \{0, 1, \dots\}$, write the set of all k -th order polynomials (complete k -th order polynomials) defined on a bounded domain $\Omega \subset \mathbb{R}^d$ as

$$\mathcal{P}_k(\Omega) = \left\{ \sum_{|\beta| \leq k} c_{\beta} x_1^{\beta_1} \cdots x_d^{\beta_d} \mid c_{\beta} \in \mathbb{R}, \beta \in \{0, 1, \dots, k\}^d, \mathbf{x} \in \Omega \right\},$$

where β is multi-indexed. In this case, the following results can be obtained (cf. [42, Theorem 14.1, p. 120], [158, Lemma 2.8, p. 60]).

Theorem 6.6.2 (Approximation Ability of the k -th Order Polynomials) *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with piecewise smooth boundary, $p \in [1, \infty]$ and $k \in \{0, 1, \dots\}$. In this case,*

$$\inf_{\phi \in \mathcal{P}_k(\Omega)} \|v - \phi\|_{W^{k+1,p}(\Omega; \mathbb{R})} \leq c |v|_{W^{k+1,p}(\Omega; \mathbb{R})} \quad (6.6.10)$$

holds with respect to an arbitrary $v \in W^{k+1,p}(\Omega; \mathbb{R})$. Here, c is a positive constant dependent on Ω and k . \square

Based on Theorem 6.6.2, the following results can be obtained with respect to the interpolation functions on the domain Ω_i of finite elements $i \in \mathcal{E}$ (cf. [42, Theorem 16.1, p. 126]).

Theorem 6.6.3 (Interpolation Error on a Finite Element) *Let $\{\mathcal{T}_h\}_{h \rightarrow 0}$ be a regular finite element division sequence with respect to $d \in \{1, 2, 3\}$ -dimensional bounded domain $\Omega \subset \mathbb{R}^d$ and $\mathcal{T}_h = \{\Omega_i\}_{i \in \mathcal{E}}$ its element. With respect to $\alpha \in \mathcal{N}_i$, let $\hat{\varphi}_{(\alpha)}$ be basis functions defined on normal coordinates on Ξ and $W(\Xi) = \text{span}(\hat{\varphi}_{(\alpha)})_{\alpha \in \mathcal{N}_i}$ the function space of an interpolation function. $p \in [1, \infty]$ and $k, l \in \{0, 1, \dots\}$ are assumed under the conditions:*

$$k + 1 > \frac{d}{p}, \quad k + 1 \geq l \quad (6.6.11)$$

to satisfy

$$\mathcal{P}_k(\Xi) \subset W(\Xi) \subset W^{l,p}(\Xi; \mathbb{R}). \quad (6.6.12)$$

Let π be an interpolation operator of Eq. (6.6.8). In this case, with respect to an arbitrary $v \in W^{k+1,p}(\Omega; \mathbb{R})$, there exists a positive constant c which does not depend on the diameter h_i of Ω_i and

$$|v - \pi v|_{W^{l,p}(\Omega_i; \mathbb{R})} \leq ch_i^{k+1-l} |v|_{W^{k+1,p}(\Omega_i; \mathbb{R})}$$

holds. \square

Using the results from Theorem 6.6.3, the following result is obtained with respect to the interpolation error on the domain Ω (cf. [42, Theorem 16.2, p. 128], [158, Theorem 2.8, p. 62] where $p = 2$ is assumed).

Corollary 6.6.4 (Interpolation Error on a Domain) *Under the assumption of Theorem 6.6.3, let $l \in \{1, 2, \dots\}$. The basis functions $\phi = (\phi_j)_{j \in \mathcal{N}}$ defined on Ω is taken to be continuous. Let π be an interpolation operator of Eq. (6.6.8). At this point, with respect to an arbitrary $v \in W^{k+1,p}(\Omega; \mathbb{R})$, there exists a positive constant c which does not depend on the maximum diameter h and*

$$|v - \pi v|_{W^{l,p}(\Omega; \mathbb{R})} \leq ch^{k+1-l} |v|_{W^{k+1,p}(\Omega; \mathbb{R})}$$

holds. \square

Estimating the error using the order with respect to the diameter of a finite element such as in Theorem 6.6.3 or Corollary 6.6.4 is called the order estimation of error.

6.6.4 Error Estimation of Finite Element Solution

The error evaluation of an interpolation function made of basis functions by the finite element method is given by the results when the exact solution in Corollary 6.6.4 is taken to be $v \in W^{k+1,p}(\Omega; \mathbb{R})$. On the other hand, Theorem 6.1.13 shows the results when measuring the basic error $u - u_h$ with $U = H^1(\Omega; \mathbb{R})$ norm of approximate solution by the Galerkin method (finite element solution). Here, the error of the finite element solution can be obtained by using the results when $p = 2$ and $l = 1$ are set in Corollary 6.6.4.

Error evaluation when measuring with $H^1(\Omega; \mathbb{R})$ norm is as follows (cf. [42, Theorem 18.1, p. 138], [158, Theorem 2.9, p. 64]).

Theorem 6.6.5 (Error Evaluation of FE Solution Due to H^1 Norm) *Let $\{\mathcal{T}_h\}_{h \rightarrow 0}$ be a regular finite element division sequence with respect to a $d \in \{1, 2, 3\}$ -dimensional polyhedron $\Omega \subset \mathbb{R}^d$ and $\mathcal{T}_h = \{\Omega_i\}_{i \in \mathcal{E}}$ its element. $p = 2, l = 1$*

and $k \in \{0, 1, \dots\}$ are to satisfy Eqs. (6.6.11) and (6.6.12). Basis functions $\phi = (\phi_j)_{j \in \mathcal{N}}$ defined on Ω are taken to be continuous. Let the set of approximate solutions in the finite element method be

$$U_h = \left\{ v_h(\bar{v}) = \bar{v} \cdot \phi \mid \bar{v} = (\bar{v}_j)_{j \in \mathcal{N}} \in \mathbb{R}^{|\mathcal{N}|} \right\}. \quad (6.6.13)$$

Let $\{u_h\}_{h \rightarrow 0}$ ($u_h \in U_h$) be a sequence of finite element solutions with respect to Problem 6.1.6 when the homogeneous fundamental boundary condition ($u_D = 0$). Here, if the exact solution is $u \in U \cap H^{k+1}(\Omega; \mathbb{R})$, there exists a positive constant c which does not depend on h and

$$\|u - u_h\|_{H^1(\Omega; \mathbb{R})} \leq ch^k |u|_{H^{k+1}(\Omega; \mathbb{R})}$$

holds. \square

Furthermore, the results when the error $u - u_h$ is measured with an $L^2(\Omega; \mathbb{R})$ norm can be obtained in the following way using a method known as the Aubin–Nitsche trick ([42, Theorem 19.2, p. 142], [158, Theorem 2.11, p. 66], where it is assumed that Ω is a two-dimensional polygonal convex domain).

Theorem 6.6.6 (Error Evaluation of FE Solution Due to L^2 Norm) *Let $\{\mathcal{T}_h\}_{h \rightarrow 0}$ be a regular finite element division sequence with respect to a $d \in \{1, 2, 3\}$ -dimensional polyhedral bounded domain Ω and $\mathcal{T}_h = \{\Omega_i\}_{i \in \mathcal{E}}$ be its element. Suppose Eq. (6.6.12) is satisfied under $p = 2$, $l = 1$, $d \leq 3$ and $k \geq 1$. Basis functions $\phi = (\phi_j)_{j \in \mathcal{N}}$ defined on Ω will be continuous. The set of approximate solutions in the finite element method U_h is Eq. (6.6.13). Let $\{u_h\}_{h \rightarrow 0}$ ($u_h \in U_h$) be a sequence of finite element solutions with respect to Problem 6.1.6 at the homogeneous fundamental boundary condition ($u_D = 0$). Here, if the exact solution is $u \in U \cap H^{k+1}(\Omega; \mathbb{R})$, there exists a positive constant c which does not depend on h and*

$$\|u - u_h\|_{L^2(\Omega; \mathbb{R})} \leq ch^{k+1} |u|_{H^{k+1}(\Omega; \mathbb{R})}$$

holds. \square

Let us conduct an error evaluation of finite element solutions using the above results. Results seen in Sect. 5.3 can be used for the regularity of exact solutions. First, if the smoothness (regularity) of the exact solution can be determined depending on the smoothness of the known functions, the following result is obtained.

Exercise 6.6.7 (Error Evaluation of FE Solution) In Problem 6.1.6, the open angle at the boundary between the Dirichlet boundary and Neumann boundary is $\alpha < \pi/2$ and other boundaries are taken to be smooth. Let $b \in L^2(\Omega; \mathbb{R})$ and $p_N = 0$. In this case, show the order evaluation of the error with respect to the finite element solution. \square

Answer When $b \in L^2(\Omega; \mathbb{R})$, from $-\Delta u = b$, with respect to the exact solution u , $|u|_{H^2(\Omega; \mathbb{R})} \leq c_0 \|b\|_{L^2(\Omega; \mathbb{R})}$ holds. Therefore, when $k = 1$ in Theorem 6.6.5 and Theorem 6.6.6,

$$\|u - u_h\|_{H^1(\Omega; \mathbb{R})} \leq c_1 h |u|_{H^2(\Omega; \mathbb{R})},$$

$$\|u - u_h\|_{L^2(\Omega; \mathbb{R})} \leq c_1 h^2 |u|_{H^2(\Omega; \mathbb{R})}$$

are obtained with respect to the finite element solution. \square

Next, let us think about a two-dimensional domain with a non-smooth boundary.

Exercise 6.6.8 (Error Evaluation for Non-smooth Boundary) In Problem 6.1.6, Ω is taken to be a two-dimensional domain with corner point x_0 such as in Fig. 5.2. Consider the error evaluation of the finite element solution around x_0 . \square

Answer A singularity appears when Γ_1 and Γ_2 have the same boundary with an opening angle of $\alpha > \pi$ (concave angle) and the opening angle at a mixed boundary is $\alpha > \pi/2$. For example, if there is a crack at boundaries of the same type ($\alpha = 2\pi$) or a straight line at a mixed boundary ($\alpha = \pi$), from Eq. (5.3.10),

$$u \in H^{3/2-\epsilon}(B(x_0, r_0) \cap \Omega; \mathbb{R})$$

holds with respect to $\epsilon > 0$ r_0 -around the corner point. On the other hand, the order estimation of error with respect to a finite element solution must satisfy $k+1 = 2 > d/p$ with respect to $d \in \{2, 3\}$ and $p = 2$ due to Eq. (6.6.11). In other words, it is not applicable unless the exact solution is in $H^2(\Omega; \mathbb{R})$. Hence, order evaluation of the error is not possible for the finite element solution around the corner points. \square

Based on Exercise 6.6.8, order estimation of error could not be made with respect to the finite element solution around singular points. However, the following result can be obtained with respect to convergence to the exact solution due to the fact that $U \cap H^{k+1}(\Omega; \mathbb{R})$ is dense at U (cf. [95, Theorem 5.4, p. 100]).

Theorem 6.6.9 (Convergence of Finite Element Solution Using H^1 Norm) *The same notation as Theorem 6.6.5 is used. Let $\{u_h\}_{h \rightarrow 0}$ ($u_h \in U_h$) be a sequence of finite element solutions with respect to Problem 6.1.6 at the homogeneous fundamental boundary condition ($u_D = 0$). Here,*

$$\lim_{h \rightarrow 0} \|u - u_h\|_{H^1(\Omega; \mathbb{R})} = 0$$

holds with respect to the exact solution $u \in U$. \square

Furthermore, with respect to the finite element solution around singularity points, finite elements for expressing the exact solution have been thought of. For example, there are methods developed to add to basis functions which can approximate a

series expansion with respect to r around singular points as seen in Sect. 5.3 (cf. [155, Chap. 8, p. 257], [159]).

6.7 Summary

In Chap. 6, the numerical solutions with respect to boundary value problems in partial differential equations were looked at in terms of the finite element method with the Galerkin method as a leading principle and error evaluation of their numerical solutions. Key points are as follows.

- (1) The Galerkin method constructs an approximate function via a linear combination of basis functions multiplied by undetermined multipliers, and by substituting these approximate functions into the weak form, changes a boundary value problem of a partial differential equation to a simultaneous linear equation relating to undetermined multipliers (Sect. 6.1).
- (2) The finite element method is a Galerkin method. Here, the finite element method divides the domain into sets of simple shaped domains, and constructs approximate functions using basis functions of low-order polynomials in each domain and continuous at the boundaries of the split domains (Sects. 6.2, 6.3, 6.4).
- (3) The isoparametric finite element method is a method for evaluating integrals on finite elements in a normal domain by mapping finite element domains to a normal domain. Here, the mappings from finite element domains to the normal domain are taken to use the same basis functions as that used for the approximate functions with respect to the solution. For the numerical integration of a rectangle on a normal domain, Gaussian quadrature is used (Sect. 6.5).
- (4) The error norms of approximate solutions from the finite element method can be suppressed with powers of the sizes of the finite elements (Sect. 6.6).

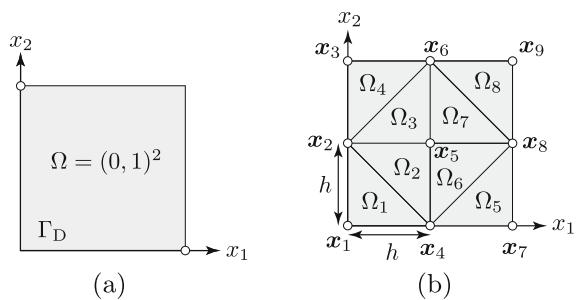
6.8 Practice Problems

6.1 Let $u : (0, 1) \rightarrow \mathbb{R}$ be the solution of the first-type boundary value problem of a one-dimensional second-order differential equation:

$$-\frac{d^2u}{dx^2} + u = 1 \quad \text{in } (0, 1), \quad u(0) = u(1) = 0.$$

Obtain the approximate solution u_h using the Galerkin method. Here, use the same basis functions as Exercise 6.1.5.

Fig. 6.40 Domain and finite element division for Exercise 6.3.2. (a) Ω and Γ_D . (b) Finite element division



6.2 Obtain the simultaneous linear equations when solving the boundary value problem in Practice 6.1 by the finite element method using the first-order basis functions. Here, let the finite element number be $m = 4$.

6.3 When the three nodes $x_{i(1)}$, $x_{i(2)}$ and $x_{i(3)}$ of the triangular finite element $i \in \mathcal{E}$ are chosen to be in the anti-clockwise direction, show that the γ defined by Eq. (6.3.8) is equal to twice the area $|\Omega_i|$ of the triangular finite element domain Ω_i .

6.4 With respect to a domain $\Omega = (0, 1)^2$ such as in Fig. 6.40a and boundaries $\Gamma_D = \{x \in \partial\Omega \mid x_1 = 0, x_2 = 0\}$ and $\Gamma_N = \partial\Omega \setminus \bar{\Gamma}_D$, obtain $u : \Omega \rightarrow \mathbb{R}$ which satisfies

$$-\Delta u = 1 \quad \text{in } \Omega, \quad \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \Gamma_N, \quad u = 0 \quad \text{on } \Gamma_D$$

via the finite element method using finite element division such as that shown in Fig. 6.40b.

6.5 In a two-dimensional Poisson problem (Problem 6.1.7 with $d = 2$), let us suppose that a rectangular first-order element of Fig. 6.26 is used. In this case, obtain the elements of the coefficient matrix and the elements of the known term vector. Here, $b = b_0$ is a constant function and $p = 0$.

6.6 When a plane stress ($\sigma_{33} = \sigma_{13} = \sigma_{23} = 0$) is assumed in a two-dimensional linear elastic problem, show the calculation method of the element coefficient matrix of a 4-node isoparametric finite element in the following order.

- Let $\bar{u}_i = (u_{11}, u_{12}, u_{13}, u_{14}, u_{21}, u_{22}, u_{23}, u_{24})^\top$ be node displacements of a finite element $i \in \mathcal{E}$. Let $\mathbf{E}(u(\xi)) = (\varepsilon_{jl}(\xi))_{jl}$ be a strain tensor and $\boldsymbol{\varepsilon}(\xi) = (\varepsilon_{11}, \varepsilon_{22}, 2\varepsilon_{12})^\top$ be its vector expression. Here, show the calculation method of displacement-strain matrix $\mathbf{B}(\xi)$ which becomes

$$\boldsymbol{\varepsilon}(\xi) = \mathbf{B}(\xi) \bar{u}_i.$$

- Let $S(\mathbf{u}(\xi)) = (\sigma_{jl}(\xi))_{jl}$ be a stress tensor and $\sigma(\xi) = (\sigma_{11}, \sigma_{22}, \sigma_{12})^\top$ its vector expression. When plane stress ($\sigma_{13} = \sigma_{23} = \sigma_{33} = 0$) is assumed, the constitutive law can be given by

$$\sigma(\xi) = D\epsilon(\xi), \quad D = \frac{e_Y}{1 - \nu_P^2} \begin{pmatrix} 1 & \nu_P & 0 \\ \nu_P & 1 & 0 \\ 0 & 0 & (1 - \nu_P)/2 \end{pmatrix}$$

using Young's modulus e_Y and Poisson ratio ν_P . Here, show the calculation method of element coefficient matrix \bar{K}_i .

Chapter 7

Abstract Optimum Design Problem



We have seen in Chaps. 5 and 6 how boundary problems of partial differential equations are constructed and how they are solved. If we compare them to the optimum design problems looked at in Chap. 1, they correspond to state determination problems. From this chapter, we finally start thinking about optimum design problems targeting the shape or topology of a domain in which the boundary value problem is defined. In this chapter, we construct an abstract problem common to both problems, and explore the ways to solve them.

In Chap. 1, the basics of the optimum design problems were looked at. In those problems considered in the chapter, the linear spaces for design variable and the state variable were both finite-dimensional vector spaces. In this chapter, such vector spaces will be extended to function spaces. Moreover, a state determination problem is replaced by an abstract variation problem. In this case the following points should be noted. The dual space of a finite-dimensional vector space was the same finite-dimensional vector space. Therefore, the derivative of the cost function with respect to a variation of the design variable is an element of the same finite-dimensional vector space as that of the design variable. However, if a function space is selected as a vector space, its dual space generally becomes a different vector space. In this chapter, there is a need to be careful of this. However, other than this, the same results as the conditions satisfied by the optimal solution in Chap. 1 should be attainable.

Furthermore, with respect to numerical solutions (algorithms), the abstract gradient method and abstract Newton method are defined by expanding the gradient method and the Newton method shown in Chap. 3. Using their solutions, the gradient method and Newton method with respect to constrained problems can be considered similarly to those shown in Chap. 3. In these cases, the numerical solutions of an abstract optimum design problem can be constructed using the algorithms shown in Chap. 3 just by replacing the corresponding terms.

Hence, if the content of this chapter is understood, the remaining issue in the shape or topology optimization problem is specifying the admissible set or function

spaces with respect to these problems within the framework of an abstract optimum design problem. In this case, clarifying the method of calculation of the Fréchet derivative with respect to a variation of design variable, and confirming that the solution with the abstract gradient method or abstract Newton method using them is included in the admissible set of the design variable become the focal points. These will be looked at for each problem in Chaps. 8 and 9.

7.1 Linear Spaces of Design Variables

Think about making the optimum design problem abstract by remembering the optimum design problem seen in Chap. 1. Here, let us use the mean compliance minimization problem (Problem 1.1.4) of the stepped one-dimensional linear elastic body in order to look at its correspondence with the abstract optimum design problem.

In Problem 1.1.4, the linear space of design variables was set to be $X = \mathbb{R}^2$ and the linear space of state variables to be $U = \mathbb{R}^2$. In this chapter, X and U can be also be function spaces. In this case, if the element of X is given, we assume that a state determination problem can be constructed, the state variables can be determined as an element of U , and cost function (functional) defined on $X \times U$ can also be calculated. Here, if solutions using the gradient or Hessian such as those looked at in Chap. 3 are considered, there is a need for X and U to be a function space in which the Fréchet derivative can be defined. Furthermore, if solutions via the abstract gradient method (Problem 7.6.1) or abstract Newton method (Problem 7.6.4) to be shown later are to be considered, X needs to be a real Hilbert space. Hence, in this chapter, we will assume X in the following way.

In the optimization problems shown in Chaps. 8 and 9, the domains over which the boundary problems of partial differential equations are defined are the scope of the design. In this case, as shown in Chap. 5, in order for the boundary problem of a partial differential equation to be defined, the domain needs to have at least a Lipschitz boundary (Sect. A.5). In Chaps. 8 and 9, the functions representing the density or domain variation are chosen to be the design variables. In this case, if the Lipschitz boundary is to be defined using these functions, the function space Y for the functions needs to be of class $C^{0,1}$. Moreover, to show the existence of the optimum solution in Sect. 7.4, Y needs to be compactly embedded in X ($\mathcal{D} \Subset X$). In fact, in Chap. 8, for the $d \in \{2, 3\}$ -dimensional bounded domain D , the function spaces X and Y will be chosen as $H^1(D; \mathbb{R})$ and $H^2(D; \mathbb{R}) \cap C^{0,1}(D; \mathbb{R})$, respectively. Similarly, in Chap. 9, for a bounded domain $D \subset \mathbb{R}^d$, X and Y will be defined as $H^1(D; \mathbb{R}^d)$ and $H^2(D; \mathbb{R}^d) \cap C^{0,1}(D; \mathbb{R}^d)$, respectively. In relation to this, $Y \Subset X$ is guaranteed by the Rellich–Kondrachov compact embedding theorem (Theorem 4.4.15). Furthermore, the admissible set \mathcal{D} of the design variables will be defined as sets satisfying additional conditions.

Hence, in this chapter, we shall denote the design variable as ϕ , which is an element of $\mathcal{D} \subset Y \Subset X$. Furthermore, since the bounded conditions correspond to

the side constraint in Chap. 1, after this chapter, we assume that ϕ is an interior point of $\mathcal{D} (\phi \in \mathcal{D}^\circ)$ when considering a gradient method or Newton method. When some of the side constraints are activated, we include them in the inequality constraints. Moreover, we will assume a variation of the design variable as $\varphi \in X$ ($\varphi \in Y$ in Chap. 9) and define the Fréchet derivatives of functions or functionals with respect to an arbitrary $\varphi \in X$ as an element in the dual space X' of X (Definition 4.4.5).

7.2 State Determination Problem

In the optimum design problem of Problem 1.1.4, the state variable was defined by \mathbf{u} and constructed so that it can be uniquely determined as a solution of the state determination problem (Problem 1.1.3) when $\mathbf{a} \in \mathcal{D}$ is given. The linear space containing \mathbf{u} was $U = \mathbb{R}^2$.

In this chapter, the state variable is written as u and uniquely determined as a solution to the state determination problem given by an abstract variational problem as shown later when the design variable $\phi \in \mathcal{D}$ is given. This problem is the same as Problem 5.2.3 but from the fact that the bilinear form a and the linear form l depend on ϕ , they are rewritten as $a(\phi)$ and $l(\phi)$ respectively. U is a real Hilbert space as per Problem 5.2.3.

Problem 7.2.1 (Abstract Variational Problem for ϕ) Let $\phi \in \mathcal{D}$ and define $a(\phi) : U \times U \rightarrow \mathbb{R}$ as a bounded and coercive bilinear form on U and $l(\phi) = l(\phi)(\cdot) = \langle l(\phi), \cdot \rangle \in U'$. In this case, find $u \in U$ such that

$$a(\phi)(u, v) = l(\phi)(v)$$

for every $v \in U$. □

Let us write Problem 7.2.1 in the following way. Let $\tau(\phi) : U \rightarrow U'$ be the isomorphism given by the Lax–Milgram theorem (Theorem 5.2.4) for a given bounded and coercive bilinear form $a(\phi)(\cdot, \cdot)$ and known term $l(\phi) \in U'$. In this case, find $u \in U$ which satisfies

$$s(\phi, u) = l(\phi) - \tau(\phi)u = 0_{U'}. \quad (7.2.1)$$

Moreover, as shown in Exercise 5.2.5 in Chap. 5, a non-homogeneous Dirichlet problem is contained in an abstract variational problem by replacing $u \in U$ in Eq. (7.2.1) by $\tilde{u} = u - u_D \in U$. Here, $l(\phi)$ can be replaced by $\hat{l}(\phi) = l(\phi) - \tau(\phi)u_D$ and becomes

$$s(\phi, \tilde{u}) = \hat{l}(\phi) - \tau(\phi)\tilde{u} = 0_{U'}. \quad (7.2.2)$$

For simplicity, in this chapter, we use Eq. (7.2.1).

Moreover, as shown in Remark 7.6.3 later, in order to define the Fréchet derivative of cost function with respect to a variation of the design variable, the solution u of Problem 7.2.1 needs to be an element of the admissible set of state variables $\mathcal{S} \subset U$. In order for this to be satisfied, the known term $l(\phi)$ or the regularity of domain need to be appropriately set. Their conditions will be shown in Chaps. 8 and 9 depending on the specific optimum design problems. Here, the design variable u is assumed to be obtained as an element of \mathcal{S} .

Under this type of setting, in a similar manner to the state determination problem (Problem 1.1.3) in Chap. 1, $v \in U$ is taken to be an adjoint variable (or a Lagrange multiplier) and

$$\mathcal{L}_S(\phi, u, v) = -a(\phi)(u, v) + l(\phi)(v) \quad (7.2.3)$$

is referred to as the Lagrange function for the state determination problem. Here, u is not necessarily the solution of Problem 7.2.1. However, the element $u \in U$ which satisfies

$$\mathcal{L}_S(\phi, u, v) = 0 \quad (7.2.4)$$

with respect to an arbitrary $v \in U$ is equivalent to the weak-form solution of Problem 7.2.1.

7.3 Abstract Optimum Design Problem

In Problem 1.1.4, the cost functions f_0 and f_1 were defined as a function of the design variable and the state variable. Here, the functionals f_0, \dots, f_m defined on the admissible set $\mathcal{D} \subset X$ of the design variables defined in Sect. 7.1 and the admissible set $\mathcal{S} \subset U$ of the state variables defined in Sect. 7.2 are set to be cost functions and used in an abstract optimum design problem as follows.

Problem 7.3.1 (Abstract Optimum Design Problem) For $(\phi, u) \in \mathcal{D} \times \mathcal{S}$, if $f_0, \dots, f_m : \mathcal{D} \times \mathcal{S} \rightarrow \mathbb{R}$ is given, obtain ϕ which satisfies

$$\min_{(\phi, u) \in \mathcal{D} \times \mathcal{S}} \{ f_0(\phi, u) \mid f_1(\phi, u) \leq 0, \dots, f_m(\phi, u) \leq 0, \text{ Problem 7.2.1} \}. \quad \square$$

Problem 7.3.1 can be thought of in the following way by using Figs. 2.1 and 2.3. Even if X becomes a real Hilbert space, there is no need to change the image of the plane within the diagrams. Furthermore, if \mathcal{D} only imposes constraint conditions such as smoothness on an element in X with no constraint conditions imposed using the norm of X directly, it again becomes a similar plane image as X . However, the plane in this case can be thought to be a plane made of only elements satisfying constraint conditions such as smoothness, just like the set of rational numbers in the real numbers. Moreover, the set S of Eq. (2.1.1) called the admissible set of design

variables in Chap. 2 can be replaced by

$$S = \left\{ (\phi, u(\phi)) \in \mathcal{D} \times \mathcal{S} \mid f_1(\phi, u(\phi)) \leq 0, \dots, f_m(\phi, u(\phi)) \leq 0, \right. \\ \left. s(\phi, u) = 0_{U'} \right\} \quad (7.3.1)$$

in this chapter. This set is an image of the sets on a plane satisfying $f_1 \leq 0$ and $f_2 \leq 0$ in Figs. 2.1 and 2.3.

We shall now look at the Fréchet derivatives of the cost functions and KKT conditions with respect to Problem 7.3.1. In this case, the notation of the Lagrange function is used in several ways. Here, in order to avoid confusion, let us summarize these relationships. Let the Lagrange function with respect to Problem 7.3.1 be

$$\mathcal{L}(\phi, u, v_0, v_1, \dots, v_m) = \mathcal{L}_0(\phi, u, v_0) + \sum_{i \in \{1, \dots, m\}} \lambda_i \mathcal{L}_i(\phi, u, v_i). \quad (7.3.2)$$

Here, $\lambda = \{\lambda_1, \dots, \lambda_m\}^\top$ is a Lagrange multiplier with respect to $f_1 \leq 0, \dots, f_m \leq 0$. Furthermore, if the cost function f_i is given as a functional of the solution u of a state determination problem (Problem 7.2.1),

$$\mathcal{L}_i(\phi, u, v_i) = f_i(\phi, u) + \mathcal{L}_S(\phi, u, v_i) \quad (7.3.3)$$

is referred to as the Lagrange function with respect to $f_i(\phi, u)$. Here, \mathcal{L}_S is the Lagrange function with respect to Problem 7.2.1 defined by Eq. (7.2.3). Moreover, v_i is a Lagrange multiplier defined with respect to f_i . If f_i contains boundary integrals on the Dirichlet boundary, for example $\int_{\Gamma_D} v_{Di} \partial_\nu u \, d\gamma$ with respect to a Poisson problem, $\tilde{v}_i = v_i - v_{Di}$ is assumed to be an element of U . Details are shown in Chaps. 8 and 9. Hereinafter, boundary integrals on the Dirichlet boundary are not included in f_i and v_i is an element of U .

7.4 Existence of an Optimum Solution

The abstract optimum design problem was defined in Problem 7.3.1. In this section, we will confirm the existence of an optimum solution. To do this, Weierstrass's theorem (Theorem 2.3.2) shown in Chap. 2 becomes a basic principle. Here, we will consider a corresponding theorem for the abstract optimum design problem. The concept used here are explained precisely in [62, Section 2.3, p. 38, Section 2.4, p. 45].

In the optimization problems considered in Chap. 2, the cost functions were defined only for the design variables. In contrast, in the case of optimum design problems, the cost functions are defined as functions of the design variable ϕ and state variable $u(\phi)$ which is determined with ϕ . Hence, the assumption of

Theorem 2.3.2 in Chap. 2 that an admissible set of design variables is a bounded closed subset is replaced with that the admissible set for $(\phi, u(\phi))$, or its subset \mathcal{S} defined in Eq. (7.3.1), is compact on $X \times U$ in the case of abstract optimum design problems. Since $u(\phi)$ is continuously determined from $\phi \in \mathcal{D}$, the admissible set of $(\phi, u(\phi))$ is given by the graph of $u(\phi)$ with respect to ϕ defined as

$$\mathcal{F} = \{(\phi, u(\phi)) \in \mathcal{D} \times \mathcal{S} \mid \text{Problem 7.2.1}\}. \quad (7.4.1)$$

Then, we need to show that \mathcal{F} is compact on $X \times U$ in the abstract optimum design problem (Problem 7.3.1). To do it, we form the following assumption in addition to the hypothesis shown in the abstract variational problem (Problem 7.2.1).

Hypothesis 7.4.1 (Continuity of $a(\phi)$ and $l(\phi)$) Let $a(\phi)$ and $l(\phi)$ defined in Problem 7.2.1 be continuous with respect to $\phi \in \mathcal{D}$, that is, $a(\phi_n) \rightarrow a(\phi)$ and $l(\phi_n) \rightarrow l(\phi)$ hold with respect to an arbitrary Cauchy sequence $\phi_n \rightarrow \phi$ on X which is uniformly convergent in \mathcal{D} . \square

The compactness of \mathcal{F} can be shown as follows[62, Lemma 2.1, p. 14, Lemma 2.12, p. 39].

Lemma 7.4.2 (Compactness of \mathcal{F}) *In addition to the hypothesis in Problem 7.2.1, let Hypothesis 7.4.1 be satisfied. With respect to an arbitrary Cauchy sequence $\phi_n \rightarrow \phi$ on X which is uniformly convergent in \mathcal{D} and the solutions $u_n = u(\phi_n) \in U$ ($n \rightarrow \infty$) of Problem 7.2.1, the convergence*

$$u_n \rightarrow u \quad \text{strongly in } U$$

holds, and $u = u(\phi) \in U$ solves Problem 7.2.1. \square

Proof With respect to the solution u_n of Problem 7.2.1 for ϕ_n ,

$$\alpha_n \|u_n\|_U^2 \leq a(\phi_n)(u_n, u_n) = l(\phi_n)(u_n) \leq \|l(\phi_n)\|_{U'} \|u_n\|_U$$

holds. Here, α_n is a positive constant used in the definition of coerciveness for $a(\phi_n)$. When $\phi_n \rightarrow \phi$ is uniformly convergent in \mathcal{D} , α_n can be replaced by a positive constant α not depending on n . From the equation, it can be confirmed that $\{u_n\}_{n \in \mathbb{N}}$ is bounded. Then, there exists a subsequence such that $u_n \rightarrow u$ weakly in U .

Next, we will show that u is the solution of Problem 7.2.1 for ϕ . From the definition of Problem 7.2.1,

$$\lim_{n \rightarrow \infty} a(\phi_n)(u_n, v) = \lim_{n \rightarrow \infty} l(\phi_n)(v) \quad (7.4.2)$$

holds with respect to an arbitrary $v \in U$. Using Hypothesis 7.4.1, the right-hand side of Eq. (7.4.2) becomes

$$\lim_{n \rightarrow \infty} l(\phi_n)(v) = l(\phi)(v). \quad (7.4.3)$$

The left-hand side of Eq. (7.4.2) becomes

$$\lim_{n \rightarrow \infty} a(\phi_n)(u_n, v) = a(\phi)(u, v). \quad (7.4.4)$$

Indeed, we have Eq. (7.4.4) from

$$\begin{aligned} & |a(\phi_n)(u_n, v) - a(\phi)(u, v)| \\ & \leq |a(\phi_n)(u_n, v) - a(\phi)(u_n, v)| + |a(\phi)(u_n, v) - a(\phi)(u, v)| \\ & \leq \|a(\phi_n) - a(\phi)\|_{\mathcal{L}(U \times U, \mathbb{R})} \|u_n\|_U \|v\|_U + |a(\phi)(u_n - u, v)| \end{aligned}$$

and by using Hypothesis 7.4.1 and $u_n \rightarrow u$ weakly in U . Substituting Eqs. (7.4.3) and (7.4.4) in Eq. (7.4.2), it is confirmed that u is the solution of Problem 7.2.1 for ϕ .

Since the weak convergence was shown, then to prove the strong convergence of $\{u_n\}_{n \in \mathbb{N}}$ to u , it is sufficient to show that

$$\|u_n\|_U \rightarrow \|u\|_U \quad (n \rightarrow \infty). \quad (7.4.5)$$

Indeed, when using

$$\|\|v\|\| = \langle \tau(\phi)v, v \rangle$$

as a norm on U , we have

$$\begin{aligned} \|\|u_n\|\| &= \langle \tau(\phi)u_n, u_n \rangle = \langle (\tau(\phi) - \tau(\phi_n))u_n, u_n \rangle + \langle \tau(\phi_n)u_n, u_n \rangle \\ &= \langle (\tau(\phi) - \tau(\phi_n))u_n, u_n \rangle + l(\phi_n)(u_n) \\ &\rightarrow l(\phi)(u) = \|\|u\|\| \quad (n \rightarrow \infty). \end{aligned} \quad (7.4.6)$$

Then, $u_n \rightarrow u$ strongly in U is proved. \square

We assume that $u(\phi)$ belongs to \mathcal{S} is guaranteed in the setting of Problem 7.2.1.

On the other hand, we form the following hypothesis for the objective function.

Hypothesis 7.4.3 (Continuity of f_0) Let f_0 be lower semi-continuous on S defined in Eq. (7.3.1). That is, with respect to an arbitrary Cauchy sequence $\phi_n \rightarrow \phi$ on X which is uniformly convergent in \mathcal{D} , by which we determine a Cauchy sequence $u(\phi_n) \rightarrow u(\phi)$ ($(\phi_n, u(\phi_n)), (\phi, u(\phi)) \in S$), it holds that

$$\lim_{n \rightarrow \infty} \inf f_0(\phi_n, u(\phi_n)) \geq f_0(\phi, u(\phi)). \quad \square$$

Using the hypotheses and the previous lemma above, we have the following result for the existence of a solution to the abstract optimum design problem (Problem 7.3.1)[62, Theorem 2.1, p. 16, Theorem 2.8, p. 41].

Theorem 7.4.4 (Existence of an Optimum Solution) *In addition to the hypothesis in Problem 7.2.1, suppose Hypothesis 7.4.1 is satisfied. Let S in Eq. (7.3.1) not be empty and compact in $X \times U$. Moreover, f_0 is lower semi-continuous (Hypothesis 7.4.3) on S . Then, there exists a minimum point in Problem 7.3.1. \square*

Proof Let $\{\phi_n\}_{n \in \mathbb{N}}$ ($\phi_n \in \mathcal{D}$) be a minimizing sequence in Problem 7.3.1, and

$$q = \inf_{(\phi, u(\phi)) \in S} f_0(\phi, u(\phi)) = \lim_{n \rightarrow \infty} f_0(\phi_n, u(\phi_n)). \quad (7.4.7)$$

Since \mathcal{D} is compact, there exists a subsequence which we still denote by $\{\phi_n\}_{n \in \mathbb{N}}$ and $\phi^* \in \mathcal{D}$ such that

$$\phi_n \rightarrow \phi^* \quad \text{strongly in } X. \quad (7.4.8)$$

From Lemma 7.4.2, we have

$$u(\phi_n) \rightarrow u(\phi^*) \quad \text{strongly in } U, \quad (7.4.9)$$

where $u(\phi_n)$ and $u(\phi^*)$ are the solutions of Problem 7.2.1 with respect to ϕ_n and ϕ^* , respectively. Using Eqs. (7.4.8), (7.4.9), (7.4.7) and the lower semi-continuity of f_0 on S , we conclude that the limit

$$q = \lim_{n \rightarrow \infty} \inf f_0(\phi_n, u(\phi_n)) = f_0(\phi^*, u(\phi^*))$$

holds. It means that $(\phi^*, u(\phi^*)) \in S$ is a minimum point in Problem 7.3.1. \square

7.5 Derivatives of Cost Functions

From this point onward, assuming that the conditions for the existence of a solution of the abstract optimum design problem (Problem 7.3.1) are satisfied, we shall examine a solution to an optimization problem with equality constraints. In this book, we focus on an approach based on the gradient method, so next let us think about the way to seek the Fréchet derivative of a cost function f_i with respect to a variation of ϕ on X . Here, there is a need to seek the Fréchet derivative on X when the equality constraints of the abstract variational problem (Problem 7.2.1) are satisfied. With respect to the equality-constrained problems on a finite-dimensional vector space, the Lagrange multiple method described in Sect. 2.6.2 (or adjoint variable method described in Sect. 2.6.5) was used. This principle is based on Theorem 2.6.4. Here, let us think about expanding this into the function space.

Let us expand Problem 2.6.1 to a problem defined on $X \times U$. Let us consider an optimization problem with equality constraint such as the following. Here, let f_i be a cost function for $i \in \{1, \dots, m\}$.

Problem 7.5.1 (Optimization Problem with Equality Constraint) Let $(\phi, u) \in X \times U$. If $f_i : X \times U \rightarrow \mathbb{R}$ is given, find (ϕ, u) which satisfies

$$\min_{(\phi, u) \in X \times U} \{ f_i(\phi, u) \mid s(\phi, u) = 0_{U'} \},$$

where $s(\phi, u)$ is defined by Eq. (7.2.1). \square

In this chapter, an arbitrary variation of (ϕ, u) will be denoted by $(\varphi, w) \in X \times U$ and the Fréchet derivative of s and f_i will be written as

$$\begin{aligned} f'_i(\phi, u)[\varphi, w] &= f_{i\varphi}(\phi, u)[\varphi] + f_{iu}(\phi, u)[w] \\ &= \langle g_{f_i}, \varphi \rangle + f_{iu}(\phi, u)[w], \end{aligned} \quad (7.5.1)$$

$$\begin{aligned} s'(\phi, u)[\varphi, w] &= s_\varphi(\phi, u)[\varphi] + s_u(\phi, u)[w] \\ &= g_h[\varphi] - \tau(\phi)w, \end{aligned} \quad (7.5.2)$$

respectively. We shall use these notations to show the result of Theorem 2.6.4 being expanded.

Theorem 7.5.2 (1st Necessary Condition for a Minimizer) Let f_i and s of Problem 7.5.1 be elements of $C^1(X \times U; \mathbb{R})$ and $C^1(X \times U; U')$, respectively. Let the Fréchet derivatives of f_i and s with respect to an arbitrary $\varphi \in X$ be given by Eqs. (7.5.1) and (7.5.2), respectively. In this case, if (ϕ, u) is the minimal point of Problem 7.5.1, there exists a $v_i \in U$ which satisfies

$$\langle g_{f_i}, \varphi \rangle + \langle g_h[\varphi], v_i \rangle + \langle f_{iu}(\phi, u) - \tau^*(\phi)v_i, w \rangle = 0, \quad (7.5.3)$$

$$\langle l(\phi) - \tau(\phi)u, w \rangle = 0 \quad (7.5.4)$$

for an arbitrary $(\varphi, w) \in X \times U$. Here, $\tau^*(\phi) : U \rightarrow \hat{U}$ is the adjoint operator of $\tau(\phi)$. \square

Proof From the fact that we assume $s \in C^1(X \times U; U')$ and there is a unique solution u which satisfies $s(\phi, u) = 0_{U'}$, s satisfies the following assumptions for the implicit function theorem (Theorem A.4.2) in neighborhood $B_X \times B_U \subset X \times U$ of $(\phi, u) \in X \times U$:

- (1) $s(\phi, u) = 0_{U'}$,
- (2) $s \in C^0(B_X \times B_U; U')$,
- (3) $s(\phi, \cdot) \in C^1(B_U; U')$ with respect to an arbitrary $y = (\varphi, w) \in B_X \times B_U$ and $s_u(\phi, u) = -\tau : U \rightarrow U'$ is continuous at (ϕ, u) ,
- (4) $(s_u(\phi, u))^{-1} = -\tau^{-1} : U' \rightarrow U$ is bounded and linear.

Hence, from the implicit function theorem, there exist some neighborhood $U_X \times U_U \subset B_X \times B_U$ and continuous mapping $v : U_X \rightarrow U_U$ (v is the Greek letter upsilon), and $s(\phi, u) = 0_{U'}$ can be written as

$$u = v(\phi). \quad (7.5.5)$$

Therefore, $y(\phi) = (\phi, v(\phi)) \in C^1(\mathcal{D}; X \times U)$ can be defined.

Hence, write $\tilde{f}_i(\phi) = f_i(\phi, v(\phi)) = f_i(y(\phi))$. Since $f_i \in C^1(X \times U; \mathbb{R})$, when ϕ is a local minimizer,

$$\tilde{f}'_i(\phi)[\varphi] = y'^*(\phi) \circ g_i(\phi, v(\phi))[\varphi] = 0 \quad (7.5.6)$$

holds with respect to an arbitrary $\varphi \in X$. Here,

$$\begin{aligned} g_i(\phi, v(\phi)) &= f'_i(\phi, v(\phi)) \in \mathcal{L}(X; X' \times U') = \mathcal{L}(X; \mathcal{L}(X \times U; \mathbb{R})), \\ y'(\phi) &\in \mathcal{L}(X; X \times U), \quad y'^*(\phi) \in \mathcal{L}(X' \times U'; X'). \end{aligned}$$

$\mathcal{L}(X; U)$ represents the bounded linear operator $X \rightarrow U$. \circ represents a composition operator. We rewrite the relationship of Eq. (7.5.6) as follows.

Firstly, let us write the admissible set of (ϕ, u) as

$$S = \{(\phi, u) \in X \times U \mid s(\phi, u) = 0_{U'}\}. \quad (7.5.7)$$

For $y(\phi) = (\phi, u) \in S$, we denote the kernel of $s'(\phi, u) \in \mathcal{L}(X \times U; U')$ by

$$T_S(\phi, u) = \{(\varphi, \hat{v}) \in X \times U \mid s'(\phi, u)[\varphi, \hat{v}] = 0_{U'}\} \quad (7.5.8)$$

and the space orthogonal to $T_S(\phi, u)$ as

$$\begin{aligned} T'_S(\phi, u) \\ = \{(\psi, w) \in X' \times U' \mid \langle (\varphi, \hat{v}), (\psi, w) \rangle = 0 \text{ for all } (\varphi, \hat{v}) \in T_S(\phi, u)\}. \end{aligned} \quad (7.5.9)$$

Moreover, the relationship between $T_S(\phi, u)$ and the Fréchet derivative $y'(\phi)$ of $y(\phi)$ can be obtained in the following way. If we take the Fréchet derivative on both sides of $s(\phi, u) = 0_{U'}$ with respect to an arbitrary $\varphi \in X$, we get

$$s'(\phi, u) \circ y'(\phi)[\varphi] = 0_{U'}. \quad (7.5.10)$$

Here, the invertibility of $\tau(\phi)$ was used. This relationship shows that the image space $\text{Im } y'(\phi)$ of $y'(\phi)$ is actually the kernel space $\text{Ker } s'(\phi, u)$ of $s'(\phi, u)$. In other words, the following is established:

$$T_S(\phi, u) = \text{Im } y'(\phi). \quad (7.5.11)$$

We use the relationship above to rewrite Eq. (7.5.6). When ϕ is a local minimizer, $g_i(\phi, v(\phi))$ needs to be orthogonal to an arbitrary $(\varphi, v_i) \in T_S(\phi, u)$. Hence,

$$g_i(\phi, v(\phi)) \in T_S'(\phi, u). \quad (7.5.12)$$

Here, from the theorem relating to the orthogonal complement space of the image space and the kernel space and Eq. (7.5.11),

$$T_S'(\phi, u) = (T_S(\phi, u))^\perp = (\text{Ker } s'(\phi, u))^\perp = \text{Im } s'^*(\phi, u)$$

is established. Here, $s'^*(\phi, u) \in \mathcal{L}(U; X' \times U')$. Therefore, Eq. (7.5.12) is equivalent that there exists some $v_i \in U$ and

$$\begin{aligned} f_{i\phi}(\phi, u)[\varphi] + f_{iu}(\phi, u)[w] + \langle s_\phi(\phi, u)[\varphi], v_i \rangle + \langle s_u(\phi, u)[w], v_i \rangle \\ = \langle g_{f_i}, \varphi \rangle + \langle g_h[\varphi], v_i \rangle + \langle f_{iu}(\phi, u) - \tau^*(\phi)v_i, w \rangle = 0 \end{aligned}$$

holds with respect to an arbitrary $(\varphi, w) \in X \times U$. In other words, Eq. (7.5.3) is established. Moreover, Eq. (7.5.4) holds if u is the solution of Eq. (7.2.1). \square

7.5.1 Adjoint Variable Method

Let us define the adjoint variable method in the following way based on Theorem 7.5.2. $v_i \in U$ is called an adjoint problem with respect to f_i , and is determined so that the second term on the left-hand side of Eq. (7.5.3) becomes zero. In other words, let it be the solution of the following problem.

Problem 7.5.3 (Adjoint Problem with Respect to f_i) When $\phi \in X$ and the solution $u \in U$ of Eq. (7.2.1) in this case as well as $f_{iu}(\phi, u) \in U'$ are given, obtain a function $v_i \in U$ which satisfies

$$f_{iu}(\phi, u) - \tau^*(\phi)v_i = 0_{U'}. \quad (7.5.13)$$

Here, $\tau(\phi)$ is the same as Eq. (7.2.1). \square

If the solution v_i of Problem 7.5.3 is used, Eq. (7.5.3) can be written as

$$\langle g_{f_i}, \varphi \rangle + \langle g_h[\varphi], v_i \rangle = \langle g_i, \varphi \rangle = 0. \quad (7.5.14)$$

When u is the solution of Eq. (7.2.1) and (ϕ, u) is the minimal point of Problem 7.5.1, Eq. (7.5.14) holds by Theorem 7.5.2.

The g_i in this case is the gradient of the Fréchet derivative of f_i with respect to $\varphi \in X$ when u continues to be the solution of the state determination problem (Problem 7.2.1), even when the design variable varies with an arbitrary $\varphi \in X$. Here,

if $v(\phi)$ of Eq. (7.5.5) defined in the proof of Theorem 7.5.2 is used, the following can be written with respect to $\tilde{f}_i(\phi) = f_i(\phi, v(\phi))$:

$$\tilde{f}'_i(\phi)[\varphi] = \langle g_i, \varphi \rangle. \quad (7.5.15)$$

7.5.2 Lagrange Multiplier Method

The gradient g_i of the Fréchet derivative of the cost function f_i with respect to an arbitrary variation $\varphi \in X$ of design variable can also be obtained from the Lagrange multiplier method shown next. This method is used in Chaps. 8 and 9 because the process is explicit.

As defined in Problem 2.6.5, the Lagrange multiplier method is a method for finding candidates for solutions by replacing optimization problems with equality constraints with stationary conditions of Lagrange functions. Hence, set the Lagrange function of Problem 7.5.1 to be

$$\mathcal{L}_i(\phi, u, v_i) = f_i(\phi, u) + \langle s(\phi, u), v_i \rangle = f_i(\phi, u) + \mathcal{L}_S(\phi, u, v_i). \quad (7.5.16)$$

Here, $\mathcal{L}_S(\phi, u, v_i)$ is a Lagrange function of the state determination problem (Problem 7.2.1). The function u is not necessary for the solution of Problem 7.2.1. v_i is the Lagrange multiplier with respect to the state determination problem prepared for f_i and assumed to be an element of U , similarly to Theorem 7.5.2. In this case, the Fréchet derivative of $\mathcal{L}_i(\phi, u, v_i)$ with respect to an arbitrary variation $(\varphi, \hat{u}, \hat{v}_i) \in X \times U \times U$ of (ϕ, u, v_i) becomes

$$\begin{aligned} \mathcal{L}'_i(\phi, u, v_i)[\varphi, \hat{u}, \hat{v}_i] \\ = \mathcal{L}_{i\phi}(\phi, u, v_i)[\varphi] + \mathcal{L}_{iu}(\phi, u, v_i)[\hat{u}] + \mathcal{L}_{iv_i}(\phi, u, v_i)[\hat{v}_i]. \end{aligned} \quad (7.5.17)$$

The following can be obtained with respect to the third term on the right-hand side of Eq. (7.5.17):

$$\mathcal{L}_{iv_i}(\phi, u, v_i)[\hat{v}_i] = \mathcal{L}_S(\phi, u, \hat{v}_i). \quad (7.5.18)$$

The right-hand side of Eq. (7.5.18) is the Lagrange function of the state determination problem (Problem 7.2.1). In this case, if u is a solution of the state determination problem, the third term on the right-hand side of Eq. (7.5.17) becomes zero. Moreover, the following equation holds:

$$\begin{aligned} \mathcal{L}_{iu}(\phi, u, v_i)[\hat{u}] &= f_{iu}(\phi, u)[\hat{u}] + \mathcal{L}_{Su}(\phi, u, v_i)[\hat{u}] \\ &= \langle f_{iu}(\phi, u) - \tau^*(\phi)v_i, \hat{u} \rangle. \end{aligned} \quad (7.5.19)$$

The conditions under which Eq. (7.5.19) becomes zero with respect to an arbitrary $\hat{u} \in U$ is the same as the weak-form equation of the adjoint problem (Problem 7.5.3). Hence, if the weak solution of the adjoint problem is set to be v_i , the second term on the right-hand side of Eq. (7.5.17) becomes zero.

Furthermore, the first term on the right-hand side of Eq. (7.5.17) becomes

$$\mathcal{L}_{i\phi}(\phi, u, v_i)[\varphi] = \langle g_{f_i}, \varphi \rangle + \langle g_h[\varphi], v_i \rangle = \langle g_i, \varphi \rangle. \quad (7.5.20)$$

The g_i of Eq. (7.5.20) matches the g_i of Eq. (7.5.14).

This relationship is an expression which is an abstract format of Eq. (1.1.37) in Chap. 1. In Chaps. 8 and 9, the stationary conditions of a Lagrange function with respect to f_i will be used to seek g_i as shown here.

7.5.3 Second-Order Fréchet Derivatives of Cost Functions

Furthermore, let us think about the second-order derivatives of the cost functions with respect to a variation of design variable based on the definition of a Fréchet derivative (Definition 4.5.4).

In Sect. 1.1.6, the second-order derivative of mean compliance with respect to a variation of design variables when a stepped one-dimensional linear elastic problem is set to be a state determination problem was sought by using the second-order derivative of the Lagrange function \mathcal{L}_0 . In this regard, it was crucial to substitute for the variation \hat{u} of the state variable u with the variation \hat{u} which satisfies the equality constraints of the state determination problem based on Theorems 2.6.6 and 2.6.7. Here, let us think about something similar with respect to an abstract optimum design problem (Problem 7.5.1) with equality constraints.

With respect to \mathcal{L}_i defined by Eq. (7.5.16), (ϕ, u) is thought of as a design variable based on the definitions in Chap. 2. In this case, the second-order Fréchet derivative of \mathcal{L}_i with respect to arbitrary variations (φ_1, \hat{u}_1) and $(\varphi_2, \hat{u}_2) \in T_S(\phi, u)$ of $(\phi, u) \in S$ becomes

$$\begin{aligned} & \mathcal{L}_{i(\phi,u)(\phi,u)}(\phi, u, v_i)[(\varphi_1, \hat{u}_1), (\varphi_2, \hat{u}_2)] \\ &= (\mathcal{L}_{i\phi}(\phi, u, v_i)[\varphi_1] + \mathcal{L}_{iu}(\phi, u, v_i)[\hat{u}_1])_\phi[\varphi_2] \\ & \quad + (\mathcal{L}_{i\phi}(\phi, u, v_i)[\varphi_1] + \mathcal{L}_{iu}(\phi, u, v_i)[\hat{u}_1])_u[\hat{u}_2] \\ &= \mathcal{L}_{i\phi\phi}(\phi, u, v_i)[\varphi_1, \varphi_2] + \mathcal{L}_{i\phi u}(\phi, u, v_i)[\hat{u}_1, \varphi_2] \\ & \quad + \mathcal{L}_{i\phi u}(\phi, u, v_i)[\varphi_1, \hat{u}_2] + \mathcal{L}_{i u u}(\phi, u, v_i)[\hat{u}_1, \hat{u}_2]. \end{aligned} \quad (7.5.21)$$

Using it, we have the following result corresponding to Theorem 2.6.6.

Theorem 7.5.4 (The Second-Order Necessary Condition) *Let f_i and s be elements of $C^2(X \times U; \mathbb{R})$ and $C^2(X \times U; U')$, respectively. If (ϕ, u) is a local minimizer of Problem 7.5.1,*

$$\mathcal{L}_{i(\phi,u)(\phi,u)}(\phi, u, v_i)[(\varphi, \hat{v}), (\varphi, \hat{v})] \geq 0 \quad (7.5.22)$$

holds with respect to an arbitrary $(\varphi, \hat{v}) \in T_S(\phi, u)$. \square

Proof In the proof of Theorem 7.5.2, the assumption $s(\phi, \cdot) \in C^1(B_U; U')$ for the implicit function theorem is replaced by $s(\phi, \cdot) \in C^2(B_U; U')$, and then using $v(\phi)$ in Eq. (7.5.5), $y(\phi) = (\phi, v(\phi)) \in C^2(\mathcal{D}; X \times U)$ is determined. From Eq. (7.5.10), we have

$$s''(\phi, u)[y'(\phi)[\varphi], y'(\phi)[\varphi]] = 0_{U'} \quad (7.5.23)$$

with respect to $y'(\phi)[\varphi] \in T_S(\phi, u)$. Hence, if (ϕ, u) is a local minimizer of Problem 7.5.1,

$$\mathcal{L}_{i(\phi,u)(\phi,u)}(\phi, u, v_i)[y'(\phi)[\varphi], y'(\phi)[\varphi]] = \tilde{f}_i''(\phi)[\varphi, \varphi] \geq 0 \quad (7.5.24)$$

holds with respect to $y'(\phi)[\varphi] \in T_S(\phi, u)$. \square

Moreover, corresponding to Theorem 2.6.7, we obtain the following result.

Theorem 7.5.5 (The Second-Order Sufficient Condition) *Under the assumptions of Theorem 7.5.4, if Eqs. (7.5.3) and (7.5.4) are satisfied at $(\phi, u, v_i) \in X \times U^2$ and Eq. (7.5.22) replacing \geq with $>$ holds, then (ϕ, u) is a local minimizer of Problem 7.5.1.* \square

Proof When $(\phi, u, v_i) \in X \times U^2$ is a stationary point of \mathcal{L}_i in S , with respect to an arbitrary point $y(\phi + \varphi) = y(\phi) + z(\varphi)$ in a neighborhood $B \subset S$ of $y(\phi) = (\phi, u)$, there exists a $\theta \in (0, 1)$ satisfying

$$\tilde{f}_i(\phi + \varphi) - \tilde{f}_i(\phi) = \frac{1}{2} \mathcal{L}_{i(\phi,u)(\phi,u)}(\phi + \theta\varphi, u(\phi + \theta\varphi), v_i)[z(\varphi), z(\varphi)]$$

for all $y(\phi) + z(\varphi) \in B$. From the assumption, since the right-hand side is greater than or equal to zero, $\tilde{f}_i(\phi) \leq \tilde{f}_i(\phi + \varphi)$ holds. \square

In view of Theorems 7.5.4 and 7.5.5, since the left-hand side of Eq. (7.5.24) is the Hessian of \tilde{f}_i with respect to an arbitrary variation $\varphi \in X$ of ϕ , we write it as

$h_i(\phi, u, v_i) \in \mathcal{L}^2(X \times X; \mathbb{R})$ (Definition 4.5.4). h_i is calculated as

$$\begin{aligned}
 h_i(\phi, u, v_i) & [\varphi_1, \varphi_2] \\
 &= (\mathcal{L}_{i\phi}(\phi, u, v_i) [\varphi_1] + \mathcal{L}_{iu}(\phi, u, v_i) [\hat{v}_1])_\phi [\varphi_2] \\
 &\quad + (\mathcal{L}_{i\phi}(\phi, u, v_i) [\varphi_1] + \mathcal{L}_{iu}(\phi, u, v_i) [\hat{v}_1])_u [\hat{v}_2] \\
 &= \mathcal{L}_{i\phi\phi}(\phi, u, v_i) [\varphi_1, \varphi_2] + \mathcal{L}_{iu\phi}(\phi, u, v_i) [\hat{v}_1, \varphi_2] \\
 &\quad + \mathcal{L}_{i\phi u}(\phi, u, v_i) [\varphi_1, \hat{v}_2] + \mathcal{L}_{iuu}(\phi, u, v_i) [\hat{v}_1, \hat{v}_2], \tag{7.5.25}
 \end{aligned}$$

where, in order that $(\varphi_j, \hat{v}_j) \in T_S(\phi, u)$ for $j \in \{1, 2\}$, $\hat{v}_j = v'(\phi) [\varphi_j]$ has to be determined using the equation

$$\mathcal{L}_{S\phi u}(\phi, u, v) [\varphi_j, \hat{v}_j] = 0 \tag{7.5.26}$$

for all $\varphi_j \in X$. These specific results of h_i are shown in Chaps. 8 and 9.

7.5.4 Second-Order Fréchet Derivative of Cost Function Using Lagrange Multiplier Method

The application of the Lagrange multiplier method in obtaining the second-order Fréchet derivative of a cost function is described as follows. Recalling the definition of the second-order Fréchet derivative (Definition 4.5.4), and that u and v_i are the solutions of the state determination and adjoint problems, respectively, we fix φ_1 and define the Lagrange function with respect to $\langle g_i, \varphi_1 \rangle$ in Eq. (7.5.20) by

$$\mathcal{L}_{li}(\phi, u, v_i, w_i, z_i) = \langle g_i, \varphi_1 \rangle + \mathcal{L}_S(\phi, u, w_i) + \mathcal{L}_{Ai}(\phi, v_i, z_i), \tag{7.5.27}$$

where \mathcal{L}_S is defined by Eq. (7.2.3). \mathcal{L}_{Ai} is the Lagrange function of the adjoint problem (Problem 7.5.3) with respect to f_i defined by

$$\mathcal{L}_{Ai}(\phi, v_i, z_i) = \mathcal{L}_{iu}(\phi, u, v_i) [z_i] = \langle f_{iu}(\phi, u) - \tau^*(\phi) v_i, z_i \rangle, \tag{7.5.28}$$

where $\mathcal{L}_{iu}(\phi, u, v_i) [z_i]$ is given by Eq. (7.5.19). $w_i \in U$ and $z_i \in U$ are the adjoint variables provided for u and v_i in g_i .

With respect to arbitrary variations $(\varphi_2, \hat{u}, \hat{v}_i, \hat{w}_i, \hat{z}_i) \in X \times U^4$ of (ϕ, u, v_i, w_i, z_i) , the derivative of \mathcal{L}_{li} is written as

$$\begin{aligned}
 \mathcal{L}'_{li}(\phi, u, v_i, w_i, z_i) & [\varphi_2, \hat{u}, \hat{v}_i, \hat{w}_i, \hat{z}_i] \\
 &= \mathcal{L}_{li\phi}(\phi, u, v_i, w_i, z_i) [\varphi_2] + \mathcal{L}_{liu}(\phi, u, v_i, w_i, z_i) [\hat{u}]
 \end{aligned}$$

$$\begin{aligned}
& + \mathcal{L}_{\text{li}v_i}(\phi, u, v_i, w_i, z_i) [\hat{v}_i] + \mathcal{L}_{\text{li}w_i}(\phi, u, v_i, w_i, z_i) [\hat{w}_i] \\
& + \mathcal{L}_{\text{li}z_i}(\phi, u, v_i, w_i, z_i) [\hat{z}_i]. \tag{7.5.29}
\end{aligned}$$

The fourth term on the right-hand side of Eq. (7.5.29) vanishes if u is the solution of the state determination problem. If v_i can be determined as the solution of the adjoint problem, the fifth term of Eq. (7.5.29) also vanishes.

The condition that the second term on the right-hand side of Eq. (7.5.29) satisfies

$$\mathcal{L}_{\text{li}u}(\phi, u, v_i, w_i, z_i) [\hat{u}] = 0 \tag{7.5.30}$$

with respect to arbitrary variation of $\hat{u} \in U$ gives the adjoint problem with respect to $\langle g_i, \varphi_1 \rangle$ to determine w_i . The condition that the third term on the right-hand side of Eq. (7.5.29) satisfies

$$\mathcal{L}_{\text{li}v_i}(\phi, u, v_i, w_i, z_i) [\hat{v}_i] = 0 \tag{7.5.31}$$

with respect to arbitrary variation of $\hat{v}_i \in U$ gives the adjoint problem with respect to $\langle g_i, \varphi_1 \rangle$ to determine z_i .

Here, u , v_i , $w_i(\varphi_1)$ and $z_i(\varphi_1)$ are assumed to be the weak solutions of Problems 7.2.1, 7.5.3, Eqs. (7.5.30) and (7.5.31), respectively. If we denote $f_i(\phi, u)$ by $\tilde{f}_i(\phi)$, then we can write

$$\begin{aligned}
\mathcal{L}_{\text{li}\phi}(\phi, u, v_i, w_i(\varphi_1), z_i(\varphi_1)) [\varphi_2] &= \tilde{f}_i''(\phi) [\varphi_1, \varphi_2] \\
&= g_{\text{Hi}}(\phi, \varphi_1) [\varphi_2]. \tag{7.5.32}
\end{aligned}$$

In this book, $g_{\text{Hi}}(\phi, \varphi_1) [\varphi_2]$ is called the Hesse gradient.

7.6 Descent Directions of Cost Functions

In Sect. 7.5, it became apparent that the first and second-order Fréchet derivative of cost functions $\tilde{f}_0, \dots, \tilde{f}_m$ with respect to a variation of design variable can be obtained. Hence, let us think about making the solution to the optimization problem shown in Chap. 3 abstract.

7.6.1 Abstract Gradient Method

Let us make the gradient method abstract. From now on, let us go with the notation in Chap. 3 to write $\tilde{f}_i(\phi)$ as $f_i(\phi)$. Here, let us assume that the Fréchet derivative of $f_i(\phi)$ can be calculated and think about obtaining the minimum point of $f_i(\phi)$.

In the gradient method on a finite-dimensional vector space seen in Sect. 3.3 (Problem 3.3.1), when $X = \mathbb{R}^d$, the bilinear form $a_X(\cdot, \cdot) = (\cdot) \cdot (A(\cdot))$ using the positive definite symmetric matrix A was an operator $X \times X \rightarrow \mathbb{R}$ which is coercive (Definition 5.2.1), bounded and symmetric. Focusing on these characteristics, an abstract gradient method such as the one below can be thought of.

Problem 7.6.1 (Abstract Gradient Method) Let $X \ni \mathcal{D}$ be a real Hilbert space. Let $a_X : X \times X \rightarrow \mathbb{R}$ be a coercive and bounded bilinear form on X . In other words, with respect to an arbitrary $\varphi, \psi \in X$, there exists some $\alpha, \beta > 0$ and

$$a_X(\varphi, \varphi) \geq \alpha \|\varphi\|_X^2, \quad |a_X(\varphi, \psi)| \leq \beta \|\varphi\|_X \|\psi\|_X \quad (7.6.1)$$

is taken to form. With respect to $f_i \in C^1(X; \mathbb{R})$ (Definition 4.5.4), let $g_i(\phi_k) \in X'$ be the Fréchet derivative at $\phi_k \in \mathcal{D}^\circ$ which is not a local minimizer. In this case, obtain a $\varphi_{gi} \in X$ which satisfies

$$a_X(\varphi_{gi}, \varphi) = -\langle g_i(\phi_k), \varphi \rangle \quad (7.6.2)$$

with respect to an arbitrary $\varphi \in X$. \square

In Problem 7.6.1, the symmetry $a_X(\varphi, \psi) = a_X(\psi, \varphi)$ of a_X was not assumed. This is because the desired result can be obtained later in Theorem 7.6.2 without using symmetry. In reality, it is possible to think of a non-symmetric example among coercive and bounded linear forms in a real Hilbert space. For example, with respect to an arbitrary $u, v \in X$ defined on $X = H_0^1(\Omega; \mathbb{R})$,

$$a(u, v) = \int_{\Omega} \left(\nabla u \cdot \nabla v + \frac{\partial u}{\partial x_1} v \right) dx$$

is a coercive and bounded bilinear form even though it is asymmetric.¹ However, when considering the numerical solution, it is desirable to assume the symmetry of a_X . Specific provisions are shown in response to the problems in Chaps. 8 and 9.

The following result is obtained with respect to Problem 7.6.1.

Theorem 7.6.2 (Abstract Gradient Method) *The solution φ_{gi} of Problem 7.6.1 exists uniquely in X and the inequality*

$$\|\varphi_{gi}\|_X \leq \frac{1}{\alpha} \|g_i(\phi_k)\|_{X'} \quad (7.6.3)$$

is established. Here, α is a positive constant used in Eq. (7.6.1). Furthermore, φ_{gi} is the descent direction of f_i at ϕ . \square

¹This a appears in the weak form of the problem in which a convective term is added to the homogeneous Poisson problem [F. Kikuchi, personal communication].

Proof The unique existence and Eq. (7.6.3) can be seen from the Lax–Milgram theorem. Furthermore, φ_{gi} satisfies Eq. (7.6.2), hence

$$\begin{aligned} f_i(\phi + \bar{\epsilon}\varphi_{gi}) - f_i(\phi) &= \bar{\epsilon}\langle g_i, \varphi_{gi} \rangle + o(|\bar{\epsilon}|) = -\bar{\epsilon}a_X(\varphi_{gi}, \varphi_{gi}) + o(|\bar{\epsilon}|) \\ &\leq -\bar{\epsilon}\alpha \|\varphi_{gi}\|_X^2 + o(|\bar{\epsilon}|) \end{aligned}$$

holds with respect to a positive constant $\bar{\epsilon}$. \square

Even among the abstract gradient methods, let us refer to the case when a function space of H^1 -class is chosen in X as the H^1 gradient method.

Theorem 7.6.2 shows that the solution of the abstract gradient method (Problem 7.6.1) φ_{gi} is in X . However, there is no guarantee that φ_{gi} is in \mathcal{D} . Hence, there is a need to note the following.

Remark 7.6.3 (Solution of Abstract Gradient Method) In order to use the solution φ_{gi} of the abstract gradient method (Problem 7.6.1) in the solution for the abstract optimum design problem (Problem 7.3.1), φ_{gi} should be obtained as an element of \mathcal{D} . The following needs to be noted to satisfy the condition:

- (1) In order for the solution φ_{gi} of the abstract gradient method (Problem 7.6.1) to be included in \mathcal{D} , g_i needs to be included within a set of functions with appropriate regularity. For this result, the known term $l(\phi)$ and boundary regularities need to be set appropriately in order for the solution u of the state determination problem (Problem 7.2.1) to be included in an appropriate set \mathcal{S} of functions. Furthermore, $f_{iu}(\phi, u)$ needs to be appropriately set so that the solution v_i of the adjoint problem (Problem 7.5.3) is in the appropriate function set \mathcal{S} . Details of these are shown in Chaps. 8 and 9.
- (2) When this is not possible (for example when there are special constraint conditions imposed on \mathcal{D}), in seeking φ_{gi} , or after it is been sought, there is a need to add extra procedures to satisfy the necessary constraint conditions under \mathcal{D} . \square

7.6.2 Abstract Newton Method

Next let us make the Newton method abstract. Here, let us assume that the Fréchet derivative $\langle g_i(\phi_k), \varphi \rangle$ of $f_i(\phi)$ and second-order Fréchet derivative $h_i(\phi_k)[\varphi_1, \varphi_2]$ can be calculated, and think about obtaining the minimum point of $f_i(\phi)$.

As seen in Sect. 3.5, the bilinear form $a_X(\cdot, \cdot) = (\cdot) \cdot (\mathbf{A}(\cdot))$ used in the gradient method was replaced by $h(\mathbf{x}_k)[\cdot, \cdot] = (\cdot) \cdot (\mathbf{H}(\mathbf{x}_k)(\cdot))$ using the Hessian matrix \mathbf{H} in the Newton method (Problem 3.5.1). In a real Hilbert space X , the abstract Newton method such as the following can be thought of.

Problem 7.6.4 (Abstract Newton Method) Let $X \ni \mathcal{D}$ be a real Hilbert space. With respect to $f_i \in C^2(X; \mathbb{R})$ (Definition 4.5.4), the gradient of the Fréchet

derivative and Hessian of f_i at a non-local minimum point $\phi_k \in \mathcal{D}^\circ$ are denoted as $g_i(\phi_k) \in X'$ and $h_i(\phi_k) \in \mathcal{L}^2(X \times X; \mathbb{R})$, respectively. Moreover, $a_X : X \times X \rightarrow \mathbb{R}$ is a coercive and bounded bilinear form on X . Here, obtain a $\varphi_{gi} \in X$ which satisfies

$$h_i(\phi_k)[\varphi_{gi}, \varphi] + a_X(\varphi_{gi}, \varphi) = -\langle g_i(\phi_k), \varphi \rangle \quad (7.6.4)$$

with respect to an arbitrary $\varphi \in X$. \square

In Problem 7.6.4, the a_X was introduced in order to compensate for the lack of coerciveness and boundedness of the left-hand side of Eq. (7.6.4) and to ensure the regularity of φ_{gi} . Even among the abstract Newton methods, the case when function space of H^1 -class is chosen in X is called the H^1 Newton method. In Problem 7.6.4, as in Theorem 3.5.2, when ϕ_k is sufficiently close to the local minimum, it is hoped that the point sequence generated by the abstract Newton method would have quadratic convergence to the local minimum point. Moreover, Remark 7.6.3 with respect to the solution to the abstract gradient method is valid here too.

Furthermore, in the case of the abstract Newton method when the second-order Fréchet derivative of $f_i(\phi)$ is given by the Hesse gradient, Problem 7.6.4 is replaced with the following problem.

Problem 7.6.5 (Abstract Newton Method Using Hesse Gradient) Under the assumption of Problem 7.6.4, the gradient of the Fréchet derivative of f_i , search vector and Hesse gradient of f_i at a non-local minimum point $\phi_k \in \mathcal{D}^\circ$ are denoted by $g_i(\phi_k) \in X'$, $\bar{\varphi}_{gi} \in X$ and $g_{Hi}(\phi_k, \bar{\varphi}_{gi}) \in X'$, respectively. Given a coercive and bounded bilinear form $a_X : X \times X \rightarrow \mathbb{R}$ on X , find a $\varphi_{gi} \in X$ which satisfies

$$a_X(\varphi_{gi}, \varphi) = -\langle (g_i(\phi_k) + g_{Hi}(\phi_k, \bar{\varphi}_{gi})), \varphi \rangle \quad (7.6.5)$$

with respect to an arbitrary $\varphi \in X$. \square

The solution φ_{gi} of Problem 7.6.5 accords with the solution of the abstract Newton method if $\bar{\varphi}_{gi} = \varphi_{gi}$.

7.7 Solution of Abstract Optimum Design Problem

Now that the abstract gradient method and abstract Newton method have been defined, let us think about the solution of the abstract optimum design problem (Problem 7.3.1) here.

7.7.1 Gradient Method for Constrained Problems

Firstly, let us bear in mind what was learned in Sect. 3.7 and think about the gradient method with respect to a constrained problem. Here, the gradients $g_0, \dots, g_m \in X'$ of the Fréchet derivatives of the cost functions f_0, \dots, f_m are assumed to be calculable using the method shown in Sect. 7.5.

Here, let us show the KKT conditions with respect to Problem 7.3.1. The content shown here is an expansion of the KKT conditions Eq. (1.1.51) to Eq. (1.1.54) with respect to Problem 1.1.4 in Chap. 1. In Problem 1.1.4, $X = \mathbb{R}^2$ and $U = \mathbb{R}^2$. In contrast, in Problem 7.3.1, X and U were assumed to be real Hilbert spaces. The Fréchet derivatives of cost functions with respect to an arbitrary variation of design variable are included in the dual space X' of X . If this relationship is remembered, the following result can be obtained.

Let the Lagrange function with respect to Problem 7.3.1 be

$$\mathcal{L}(\phi, \lambda) = f_0(\phi) + \sum_{i \in \{1, \dots, m\}} \lambda_i f_i(\phi). \quad (7.7.1)$$

Here, $\lambda = (\lambda_1, \dots, \lambda_m)^\top \in \mathbb{R}^m$ is a Lagrange multiplier with respect to $f_1(\phi) \leq 0, \dots, f_m(\phi) \leq 0$.

In this case, the KKT conditions with respect to Problem 7.3.1 are given by

$$g_0(\phi) + \sum_{i \in \{1, \dots, m\}} \lambda_i g_i(\phi) = 0_{X'}, \quad (7.7.2)$$

$$f_i(\phi) \leq 0 \quad \text{for } i \in \{1, \dots, m\}, \quad (7.7.3)$$

$$\lambda_i f_i(\phi) = 0 \quad \text{for } i \in \{1, \dots, m\}, \quad (7.7.4)$$

$$\lambda_i \geq 0 \quad \text{for } i \in \{1, \dots, m\}. \quad (7.7.5)$$

Let us think about the solution to the abstract optimum design problem (Problem 7.3.1) based on this condition following the gradient method with respect to the constrained problem shown in Sect. 3.7. With respect to $k \in \{0, 1, 2, \dots\}$, the trial point ϕ_k is assumed to be an element of admissible set S defined by Eq. (7.3.1). Denote the set of suffixes with respect to active constraints with respect to ϕ_k as

$$I_A(\phi_k) = \{i \in \{1, \dots, m\} \mid f_i(\phi_k) \geq 0\} = \{i_1, \dots, i_{|I_A|}\}. \quad (7.7.6)$$

If there is no confusion, denote $I_A(\phi_k)$ as I_A . Moreover, the size of the search vector (step size) is adjusted by the size of a positive constant c_a . In this case, the problem seeking the search vector $\varphi_g \in X$ satisfying the inequalities constraint with respect to the cost functions around ϕ_k is constructed in the following way.

Problem 7.7.1 (Gradient Method for Constrained Problems) Suppose that for a trial point $\phi_k \in \mathcal{D}$ of Problem 7.3.1 satisfying the inequality constraints, $f_0(\phi_k)$,

$f_{i_1}(\phi_k) = 0, \dots, f_{i_{|I_A|}}(\phi_k) = 0$ and $g_0(\phi_k), g_{i_1}(\phi_k), \dots, g_{i_{|I_A|}}(\phi_k) \in X'$ are given. Let $a_X : X \times X \rightarrow \mathbb{R}$ be a coercive and bounded bilinear form on X . Moreover, c_a is taken to be a positive constant. Obtain $\phi_{k+1} = \phi_k + \varphi_g$ which satisfies

$$q(\varphi_g) = \min_{\varphi \in X} \left\{ q(\varphi) = \frac{c_a}{2} a_X(\varphi, \varphi) + \langle g_0(\phi_k), \varphi \rangle \mid f_i(\phi_k) + \langle g_i(\phi_k), \varphi \rangle \leq 0 \text{ for } i \in I_A(\phi_k) \right\}. \quad \square$$

Similarly to Problem 3.7.1, Problem 7.7.1 is a convex optimization problem. In this regard, φ_g satisfying the KKT conditions becomes the local minimizer of Problem 7.7.1. Let us consider the solution for Problem 7.3.1 by focusing on this. The method below is an abstract version of the method shown in Sect. 3.7.

Let the Lagrange function of Problem 7.7.1 be

$$\mathcal{L}_Q(\varphi_g, \lambda) = q(\varphi_g) + \sum_{i \in I_A(\phi_k)} \lambda_i (f_i(\phi_k) + \langle g_i(\phi_k), \varphi_g \rangle).$$

Here, $\lambda = (\lambda_1, \dots, \lambda_m)^\top \in \mathbb{R}^m$ is a Lagrange multiplier with respect to the inequality constraint conditions. The KKT conditions with respect to the minimum point φ_g of Problem 7.7.1 are that the following hold with respect to an arbitrary $\psi \in X$:

$$c_a a_X(\varphi_g, \psi) + \langle g_0(\phi_k), \psi \rangle + \sum_{i \in I_A(\phi_k)} \lambda_i \langle g_i(\phi_k), \psi \rangle = 0, \quad (7.7.7)$$

$$f_i(\phi_k) + \langle g_i(\phi_k), \varphi_g \rangle \leq 0 \quad \text{for } i \in I_A(\phi_k) \quad (7.7.8)$$

$$\lambda_{k+1i} (f_i(\phi_k) + \langle g_i(\phi_k), \psi \rangle) = 0 \quad \text{for } i \in I_A(\phi_k), \quad (7.7.9)$$

$$\lambda_{k+1i} \geq 0 \quad \text{for } i \in I_A(\phi_k). \quad (7.7.10)$$

$(\varphi_g, \lambda_{k+1}) \in X \times \mathbb{R}^{|I_A|}$ satisfying these can be obtained as follows.

Let $\varphi_{g0}, \varphi_{i_1}, \dots, \varphi_{i_{|I_A|}}$ be the solution of the abstract gradient method (Problem 7.6.1). Here, Eq. (7.6.2) is changed to

$$c_a a_X(\varphi_{gi}, \psi) = -\langle g_i, \psi \rangle \quad (7.7.11)$$

with respect to an arbitrary $\psi \in X$. In this case,

$$\varphi_g = \varphi_g(\lambda_{k+1i}) = \varphi_{g0} + \sum_{i \in I_A(\phi_k)} \lambda_{k+1i} \varphi_{gi} \quad (7.7.12)$$

satisfies Eq. (7.7.7). On the other hand, Eq. (7.7.8) becomes

$$\begin{aligned} & \begin{pmatrix} \langle g_{i_1}, \varphi_{g i_1} \rangle & \cdots & \langle g_{i_1}, \varphi_{g i_{|I_A|}} \rangle \\ \vdots & \ddots & \vdots \\ \langle g_{i_{|I_A|}}, \varphi_{g i_1} \rangle & \cdots & \langle g_{i_{|I_A|}}, \varphi_{g i_{|I_A|}} \rangle \end{pmatrix} \begin{pmatrix} \lambda_{k+1 i_1} \\ \vdots \\ \lambda_{k+1 i_{|I_A|}} \end{pmatrix} \\ &= - \begin{pmatrix} f_{i_1} + \langle g_{i_1}, \varphi_{g 0} \rangle \\ \vdots \\ f_{i_{|I_A|}} + \langle g_{i_{|I_A|}}, \varphi_{g 0} \rangle \end{pmatrix}. \end{aligned}$$

This equation is written as

$$(\langle g_i, \varphi_{gj} \rangle)_{(i,j) \in I_A^2} (\lambda_{k+1 j})_{j \in I_A} = - (f_i + \langle g_i, \varphi_{g 0} \rangle)_{i \in I_A}. \quad (7.7.13)$$

In Eq. (7.7.13), the matrix $(\langle g_i, \varphi_{gj} \rangle)_{(i,j) \in I_A^2}$ is symmetric because $\langle g_i, \varphi_{gj} \rangle = a_X(\varphi_{gi}, \varphi_{gj})$. If g_1, \dots, g_m are linearly independent, Eq. (7.7.13) is solvable about λ_{k+1} . Moreover, if the active constraint functions $f_{i_1}, \dots, f_{i_{|I_A|}}$ all have the value zero, from the fact that they hold even when an arbitrary real number is multiplied all of $\varphi_{gi_1}, \dots, \varphi_{gi_{|I_A|}}$, it becomes possible to obtain λ_{k+1} even when the step size $\|\varphi_g\|_X$ is not appropriately set. In addition, as adapted in Chap. 3, Eq. (7.7.13) is solved possibly several times, removing each time the constraints where the associated Lagrange multiplier is negative (active set method), although if we set appropriate constraint functions which have trade-off property with respect to an objective function, the Lagrange multipliers become always positive.

Using the definitions so far, a simple Algorithm 3.6 shown in Sect. 3.7.1 can be applied. In this case, the following changes can be made:

- (1) Replace the design variable x and its variation y as ϕ and φ respectively.
- (2) Replace Eq. (3.7.10) with Eq. (7.7.11).
- (3) Replace Eq. (3.7.11) with Eq. (7.7.12).
- (4) Replace Eq. (3.7.12) with Eq. (7.7.13).

Furthermore, when considering a complicated Algorithm 3.7 with parameter adjustments shown in Sect. 3.7.2, the following needs to be taken into consideration:

- (i) Functionality for determining c_a , so that the initial step size becomes $\|\varphi_g\| = \epsilon_g$ with a given value ϵ_g .
- (ii) When the design variable is updated to ϕ_{k+1} , the functionality of amending $\lambda_{k+1} = (\lambda_{k+1 i})_{i \in I_A(\phi_{k+1})}$ so that $|f_i(\phi_{k+1})| \leq \epsilon_i$ and $\lambda_{k+1 i} > 0$ with respect to $i \in I_A(\phi_{k+1})$ are satisfied.

- (iii) The functionality to make suitable the admissible values $\epsilon_1, \dots, \epsilon_m$ of the constraint functions f_1, \dots, f_m with respect to the convergence determination value ϵ_0 of the objective function f_0 .
- (iv) Functionality for adjusting the step size $\|\varphi_g\|$ so that global convergence is guaranteed.

With respect to the aforementioned (i), the content shown in Sect. 3.7.2 will hold as it is by replacing \mathbf{y} by φ .

The same is true for (ii) above. In other words, Algorithm 3.7 can be used exactly by replacing the update of λ_{k+1} using Eq. (3.7.21) of the Newton–Raphson method by

$$(\delta\lambda_j)_{j \in I_A} = -((g_i(\lambda_{k+1})_I, \varphi_{gj}(\lambda_{k+1}))_{(i,j) \in I_A^2}^{-1} (f_i(\lambda_{k+1}))_{i \in I_A}). \quad (7.7.14)$$

Moreover, with respect to (iii) above, the method for replacing ϵ_i so that Eq. (3.7.25) is satisfied has already been incorporated into Algorithm 3.7.

Furthermore, with respect to (iv) above, the following type of replacement would allow Algorithm 3.7 to be used as it is. The Lagrange function with respect to the abstract optimum design problem (Problem 7.3.1) is given by $\mathcal{L}(\phi, \lambda)$ of Eq. (7.7.1). In this case, the Armijo criterion becomes

$$\begin{aligned} & \mathcal{L}(\phi_k + \varphi_g, \lambda_{k+1}) - \mathcal{L}(\phi_k, \lambda_k) \\ & \leq \xi \left\langle g_0(\phi_k) + \sum_{i \in I_A(\phi_k)} \lambda_{ki} g_i(\phi_k), \varphi_g \right\rangle \end{aligned} \quad (7.7.15)$$

with respect to a $\xi \in (0, 1)$. The Wolfe criterion is given with respect to a μ ($0 < \xi < \mu < 1$) by

$$\begin{aligned} & \mu \left\langle g_0(\phi_k) + \sum_{i \in I_A(\phi_k)} \lambda_{ki} g_i(\phi_k), \varphi_g \right\rangle \\ & \leq \left\langle g_0(\phi_k + \varphi_g) + \sum_{i \in I_A(\phi_{k+1})} \lambda_{k+1} i g_i(\phi_k + \varphi_g), \varphi_g \right\rangle. \end{aligned} \quad (7.7.16)$$

Using these replacements, Algorithm 3.7 can be applied with respect to the abstract optimum design problem (Problem 7.3.1). In this case, the following changes are made in addition to (1) to (4) above:

- (5) Replace the Armijo criterion Eq. (3.7.26) with Eq. (7.7.15).
- (6) Replace the Wolfe criterion Eq. (3.7.27) with Eq. (7.7.16).
- (7) Replace Eq. (3.7.21) by Eq. (7.7.14) to update λ by the Newton–Raphson method.

In order for this algorithm to function well, there is a need for the points made in Remark 7.6.3 to be satisfied. If these are not satisfied, there is a possibility of numerical instability arising. In order to prevent these situations, there is a need to ensure that the new design variable is always included within the admissible set \mathcal{D} by adding an appropriate process after the design variable is updated.

7.7.2 Newton Method for Constrained Problems

If the second-order Fréchet derivatives of the cost functions can be obtained, it is possible to change the gradient method with respect to a constrained problem to a Newton method with respect to a constrained problem. Here, let us use the abstract Newton method (Problem 7.6.4) in order to make Problem 3.8.1 in Chap. 3 abstract.

Problem 7.7.2 (Newton Method for Constrained Problems) At a trial point $\phi_k \in \mathcal{D}$ of Problem 7.3.1 satisfying the inequality constraints, the Lagrange multiplier $\lambda_k \in \mathbb{R}^{|I_A|}$ is assumed to satisfy Eq. (7.7.8) to Eq. (7.7.10) (where $k+1$ is viewed as k). Moreover, $f_0(\phi_k), f_{i_1}(\phi_k) = 0, \dots, f_{i_{|I_A|}}(\phi_k) = 0$ and $g_0(\phi_k), g_{i_1}(\phi_k), \dots, g_{i_{|I_A|}}(\phi_k) \in X'$ as well as $h_0(\phi_k), h_{i_1}(\phi_k), \dots, h_{i_{|I_A|}}(\phi_k) \in \mathcal{L}^2(X \times X; \mathbb{R})$ are taken to be known and

$$h_{\mathcal{L}}(\phi_k) = h_0(\phi_k) + \sum_{i \in I_A(\phi_k)} \lambda_{ik} h_i(\phi_k). \quad (7.7.17)$$

Moreover, let $a_X : X \times X \rightarrow \mathbb{R}$ be a coercive and bounded bilinear form on X . In this case, obtain $\phi_{k+1} = \phi_k + \varphi_g$ which satisfies

$$\begin{aligned} q(\varphi_g) = \min_{\varphi \in X} \left\{ q(\varphi) = \frac{1}{2} (h_{\mathcal{L}}(\phi_k)[\varphi, \varphi] + a_X(\varphi, \varphi)) + \langle g_0(\phi_k), \varphi \rangle \right. \\ \left. + f_0(\phi_k) \mid f_i(\phi_k) + \langle g_i(\phi_k), \varphi \rangle \leq 0 \text{ for } i \in I_A(\phi_k) \right\}. \quad \square \end{aligned}$$

In Problem 7.7.2, the a_X was introduced in order to compensate for the lack of coerciveness and boundedness of $h_{\mathcal{L}}(\phi_k)$ on X and to ensure the regularity of φ_{gi} .

Problem 7.7.2 is classified to be a second-order optimization problem. When $h_{\mathcal{L}}(\phi_k)[\varphi, \varphi] + a_X(\varphi, \varphi)$ is a coercive and bounded bilinear form on X , Problem 7.7.2 becomes a convex optimization problem. It is not necessarily the case. However, a φ_g satisfying the KKT conditions shown next is a candidate for the minimum point with respect to Problem 7.7.2. Focusing on this, let us look at what has been learned in Sect. 3.8 in order to think of the solution to Problem 7.7.2.

It is assumed that KKT conditions at the minimum point φ_g of Problem 7.7.2 hold. In other words, the following holds with respect to an arbitrary $\psi \in X$:

$$h_{\mathcal{L}}(\phi_k)[\varphi, \psi] + a_X(\varphi, \psi) + \langle g_0(\phi_k), \psi \rangle$$

$$+ \sum_{i \in I_A(\phi_k)} \lambda_{k+1i} \langle g_i(\phi_k), \psi \rangle = 0, \quad (7.7.18)$$

$$f_i(\phi_{k+1}) = f_i(\phi_k) + \langle g_i(\phi_k), \varphi_g \rangle \leq 0 \quad \text{for } i \in I_A(\phi_k), \quad (7.7.19)$$

$$\lambda_{k+1i} (f_i(\phi_k) + \langle g_i(\phi_k), \varphi_g \rangle) = 0 \quad \text{for } i \in I_A(\phi_k), \quad (7.7.20)$$

$$\lambda_{k+1i} \geq 0 \quad \text{for } i \in I_A(\phi_k). \quad (7.7.21)$$

$(\varphi_g, \lambda_{k+1}) \in X \times \mathbb{R}^{|I_A|}$ satisfying these can be obtained as follows.

In the gradient method (Sect. 7.7.1) with constraints, $\varphi_{g0}, \varphi_{i1}, \dots, \varphi_{i|I_A|}$ were taken to be the solution for the abstract gradient method. Here, these are replaced by the solution of the abstract Newton method. Problem 7.6.4 is rewritten as follows. “Let the known functions in Problem 7.7.2 be given. Find $\varphi_{gi} \in X$ which satisfy the following with respect to an arbitrary $\varphi \in X$:

$$h_{\mathcal{L}}(\phi_k)[\varphi_{gi}, \varphi] + a_X(\varphi_{gi}, \varphi) = -\langle g_i(\phi_k), \varphi \rangle. \quad (7.7.22)$$

Here, φ_g defined by Eq. (7.7.12) satisfies Eq. (7.7.18). On the other hand, Eq. (7.7.19) becomes Eq. (7.7.13). Hence, if Eq. (7.7.13) is used to obtain λ_{k+1} , Eq. (7.7.19) is established and the KKT conditions at the minimal point φ_g of Problem 7.7.2 hold. In this case, Eqs. (7.7.20) and (7.7.21) are satisfied by choosing $I_A(\phi_{k+1})$ appropriately in the algorithm using the active set method.

Using the definitions so far, Algorithm 3.8 shown in Sect. 3.8.1 can be applied. In this case, it will be changed in the following way:

- (1) Replace the design variable x and its fluctuation y by ϕ and φ respectively.
- (2) Replace Eq. (3.8.9) with Eq. (7.7.22).
- (3) Replace Eq. (3.7.11) with Eq. (7.7.12).
- (4) Replace Eq. (3.7.12) with Eq. (7.7.13).

Furthermore, the abstract Newton method when the second-order Fréchet derivative of $f_i(\phi)$ is obtained as a Hesse gradient can be illustrated as follows. Equations (7.7.17) and (7.7.22) are replaced with

$$g_{H\mathcal{L}}(\phi_k, \bar{\varphi}_g) = g_{H0}(\phi_k, \bar{\varphi}_g) + \sum_{i \in I_A(\phi_k)} \lambda_{ik} g_{Hi}(\phi_k, \bar{\varphi}_g), \quad (7.7.23)$$

$$a_X(\varphi_{gi}, \varphi) = -\langle (g_i(\phi_k) + g_{H\mathcal{L}}(\phi_k, \bar{\varphi}_g)), \varphi \rangle, \quad (7.7.24)$$

respectively. Using the definitions, the following step is added:

- (5) Replace Eq. (3.8.11) with Eq. (7.7.24).

In this way, the difference between the gradient method with respect to a constrained problem and the Newton method is just that $a_X(\cdot, \cdot)$ of the abstract gradient method is replaced with $h_i(\phi_k)[\cdot, \cdot] + a_X(\cdot, \cdot)$ or g_i is replaced with $g_i(\phi_k) + g_{Hi}(\phi_k, \bar{\varphi}_g)$. However, with this method, a second-order derivative of a cost function is used. Hence, it is hoped that the characteristics of the Newton method mentioned in Remark 3.5.3 will hold. However, as explained in Remark 3.8.2, because the constraint condition is approximated to be up to first-order derivative, there is a need to be careful with the step size when the non-linearity of the constraint functions is strong.

Furthermore, with respect to the methods for achieving coerciveness of the bilinear form and adding the functionality for adjusting the step size, the explanation provided in Sect. 3.8.2 is still valid here.

Whether such a Newton method can be used or not depends on whether the calculation $h_i(\phi_k)[\cdot, \cdot]$ or $g_{Hi}(\phi_k, \bar{\varphi}_g)$ is possible or not. Let us look at the specific calculation methods of these in Chaps. 8 and 9.

7.8 Summary

In Chap. 7, abstract problems, which may be common to the optimum design problems targeting the topology and shape with respect to domain of a boundary value problem of partial differential equation such as shown in Chaps. 8 and 9, were constructed and their solutions were looked at. The following are the key points:

- (1) Real Hilbert spaces are chosen for the linear spaces of a design variable and state variable (Sect. 7.1). This is because the abstract gradient method is defined on a Hilbert space.
- (2) An abstract optimum design problem was defined with an abstract variational problem as a state determination problem (Sect. 7.2) and using cost functions defined via functionals of the design variable (Sect. 7.3) and the state variable (solution of state determination problem).
- (3) With respect to an abstract optimum design problem, the derivative of a cost function can be obtained via the adjoint variable method (Sect. 7.5.1) or the Lagrange multiplier method (Sect. 7.5.2). Moreover, the second-order derivative of a cost function can be obtained by substituting the derivative of the solution of the state determination problem when the equality constraints of the state determination problem are satisfied in the second-order derivative of the Lagrange function (Sect. 7.5.3).
- (4) The abstract gradient method is defined on a real Hilbert space (Sect. 7.6.1). The unique existence of the solution of the abstract gradient method is shown by the Lax–Milgram theorem. Moreover, the solution is on the downward slope of the cost function (Theorem 7.6.2). Furthermore, the abstract Newton method is defined as a method in which the bilinear form in the abstract gradient method is replaced by the sum of a second-order derivative of the cost function and a

bilinear form which compensates for the coerciveness and boundedness of the second-order derivative (Sect. 7.6.2).

- (5) The solution to the abstract optimum design problem is constructed with the same framework as the gradient method and Newton method with respect to constrained problems shown in Chap. 3 (Sects. 7.7.1 and 7.7.2).

Let us mention a few books which are useful references for this chapter. Chapter 5 of [78] is useful in relation to the Lagrange multiplier method on a function space. The paper [34] and Section 4.4 of [134] are useful with respect to the gradient method on function spaces.

Chapter 8

Topology Optimization Problems of Density Variation Type



From this chapter, we finally examine shape optimization problems in continua. Firstly, let us think about a problem seeking the appropriate arrangement of holes in a domain where the boundary value problem of a partial differential equation is defined. Such a problem is known as the topology optimization problem. Here, the term topology refers to the study of geometrical properties and spatial relation of objects unaffected by the continuous change of their shape or size. In mathematics, two mathematical objects are said to belong to the same topology if they are images of two homotopic maps; that is, if one can be continuously deformed into the other. Therefore, letting n be a natural number, a set of n -connected domains are regarded as belonging to the same homotopy groups. Here, the term “topology optimization” in the topology optimization problem refers to the determination of the connectivity of the design domain that optimizes an object’s material distribution through insertion and arrangement of holes in its structure. However, as will be explained in detail later, the shape of the holes actually becomes the target. Therefore, the problems dealt with in this chapter also become included in shape optimization problems in a wider sense. In this book, it will be referred to as the topology optimization problem in the sense that topology is also in the scope of the design.

Until now, various problem formulations with respect to the topology optimization problem and solutions to these problems have been proposed. By setting up a fixed domain in which a design target is included, a way to choose the characteristic function (an L^∞ -class function which takes the value 0 in holes and 1 in a domain) of the domain as the design variable was considered. In such a problem, we determine a state determination problem as a boundary value problem multiplying the characteristic function with the coefficient of the partial differential equation,

Electronic Supplementary Material The online version of this article (https://doi.org/10.1007/978-981-15-7618-8_8) contains supplementary material, which is available to authorized users.

and define the cost function by the characteristic function and the solutions of the boundary value problem. However, it was shown that such a problem does not always have a solution [123]. The basic reason for this is that a set of L^∞ -class functions does not induce enough regularity to keep the compactness of the admissible set of design variables [3]. Subsequently, the idea of applying a method (homogenization technique) for describing a continuum problem with a microstructure was shown [7, 100–103, 110, 111]. In particular, with respect to a $d \in \{2, 3\}$ -dimensional linear elastic body, the material constructed of microlayers of d types crossing over one another such as that in Fig. 8.1 are referred to as rank d material. Since the homogenized material constant of the rank d material could be analytically obtained, it was investigated in a lot of papers, see, for instance, [24, 26, 99, 128]. As a result, it was shown that in a uniform stress field, if each layer density in the direction of each principal stress is determined in proportion to the value of each principal stress, the mean compliance will be minimized as the volume is constrained [99]. However, no results could be obtained which recognize macro holes. Moreover, since rank d material has no rigidity with respect to shearing deformation, there was an issue with it not being applicable as a mechanical structure.

The generation of holes was confirmed when a continuum with a rectangular hole in a microcell Y similar to the one in Fig. 8.2 was assumed [25, 45, 112, 156]. This problem is constructed as a function optimization problem with the design variable as $(a_1, a_2, \theta)^\top : D \rightarrow \mathbb{R}^3$ in Fig. 8.2. The numerical solution of this problem was obtained by an iterative method satisfying the optimality criteria [156]. If such a numerical solution is obtained, we can define the density by the ratio of

Fig. 8.1 Rank 2 material

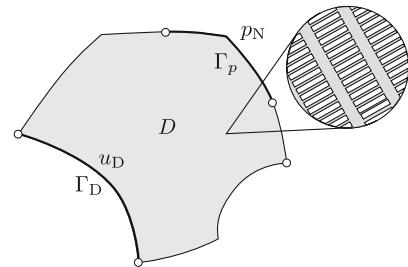
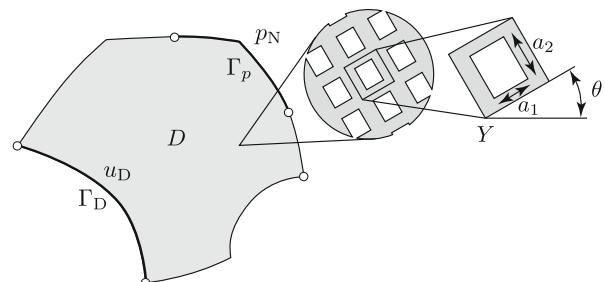


Fig. 8.2 A two-dimensional continuum with micro rectangular holes



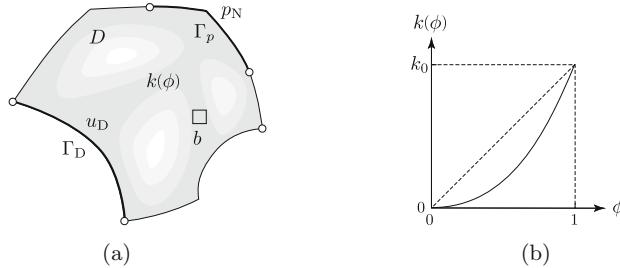


Fig. 8.3 SIMP model. (a) Boundary value problem. (b) Density ϕ and material constant $k(\phi)$

the magnitude of holes to the microcell and determine the shapes of holes from the isosurface of the density with an appropriate threshold.

After that, it was shown that even if a microstructure is not assumed, when function $\phi : D \rightarrow [0, 1]$ (in reality, in order to avoid the discontinuity of the solution of state determination problem occurring at $\phi \rightarrow 0$, $\phi : D \rightarrow [c, 1]$ is used with some small constant c) representing density is set to be the design variable, the same topology can be obtained as per the case of micro rectangular holes [117, 182]. In this case, a material characteristic (a coefficient of a partial differential equation) k was assumed to be given by

$$k(\phi) = k_0\phi^\alpha,$$

where k_0 is the existing material constant and $\alpha > 1$ is a constant. Figure 8.3 shows the image of a topology optimization problem in that case. Figure 8.3a shows an example of a state determination problem (boundary value problem). Figure 8.3b shows the relationship between density ϕ and material characteristic k . If the density which is the design variable ϕ (function defined on D) varies, the material characteristic $k(\phi)$ (function defined on D) varies via the function in Fig. 8.3b and causes the solution of the boundary value problem of Fig. 8.3a to vary. A topology optimization problem defined in this way is referred to as topology optimization problem of density type. Moreover, this problem is also referred to as the SIMP (solid isotropic material with penalization) model [138]. The reason for this can be explained in the following way. With respect to mid-level density ($\phi = 0.5$), for materials with homogeneous material such as in Fig. 8.4a, the material characteristic k becomes 0.5^α . But if, as in Fig. 8.4b, it splits into $\phi = 0$ and $\phi = 1$, k becomes 0.5. For $0.5^\alpha < 0.5$, it becomes a model which gives a penalty to materials split into $\phi = 0$ and $\phi = 1$ as having a greater material characteristic value compared with uniform materials.

With respect to the topology optimization problem of density type, the Fréchet derivatives of cost functions are calculated via the finite element method using an evaluation method such as one shown later. However, if a constant density is assumed for each finite element and moved in the negative direction of the Fréchet derivative, it has been pointed out that the density is seen to vibrate in a checkerboard



Fig. 8.4 Microstructures when density is 0.5. (a) Uniform material. (b) Material split into 0 and 1

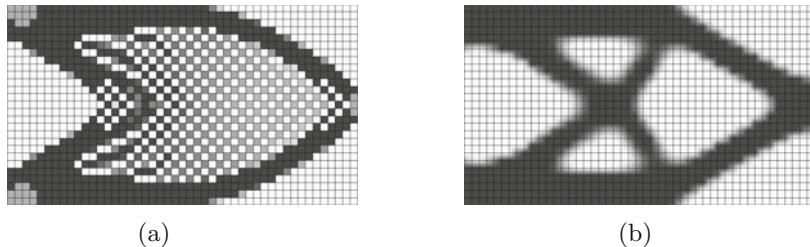


Fig. 8.5 Numerical example of density-type topology optimization problem for linear elastic body (provided by Quint Corporation). (a) Checkerboard pattern. (b) Optimal density by H^1 gradient method

pattern [46]. Figure 8.5a shows the results of numerical analysis with respect to a mean compliance minimization problem (Problem 8.9.3) of a two-dimensional linear elastic body. In a state determination problem, an external force pointing in the downward direction acts on the center of the right edge while the left edge is fixed. The black and white elements show that ϕ is close to 1 and 0, respectively. When this type of phenomenon occurs, it can be avoided by implementing post-processing such as filtering, etc. [94, 107, 151]. Moreover, methods to approximate the distribution of material parameters in a microstructure with continuous basis functions were shown [113]. However, it has been pointed out that a different issue called the island phenomenon, etc., may arise [137]. Moreover, to remove the intermediate density regions, methods using projection from intermediate density to zero one values were proposed [150, 168].

In contrast, in this chapter, we shall think about the numerical analysis of density-type topology optimization problem along the framework of abstract optimization design problem shown in Chap. 7. However, in this chapter, a function θ defined on D is newly chosen as the design variable rather than choosing directly the density to be the design variable, for reasons shown later. Hence, the density-type topology optimization problem in such a scenario will be referred to as the topology optimization problem of θ -type. The framework of the logic used in this chapter is shown clearly in the paper [11]. Figure 8.5b is the result obtained from the algorithm shown in Sect. 8.7. The fact that no numerical instability phenomenon such as the checkerboard pattern is generated can be seen.

This chapter is constructed in the following way. In Sect. 8.1, the admissible set of the design variable θ is defined in order to construct a topology optimization problem in continuum. In Sect. 8.2, a θ -type Poisson problem is defined as a state determination problem assuming that a design variable is given. Using the design variable and the solution (state variables) of the state determination problem, the topology optimization problem of θ -type will be defined in Sect. 8.3. Here, a cost function of general form will be used. The existence of a solution to the topology optimization problem of θ -type is shown in Sect. 8.4. In Sect. 8.5, by referring to the Fréchet derivatives of cost functions with respect to variation of the design variable θ as θ -derivatives, the process for obtaining the θ -derivatives and second-order θ -derivatives of cost functions will be seen based on the methods for seeking Fréchet derivatives of cost functions shown in Sect. 7.5. As a result, depending on the setting of the state determination problem, it becomes apparent that the θ -derivatives of the cost functions do not have regularity such that it would be included in the admissible set of design variables. In Sect. 8.6, the abstract gradient method and abstract Newton method are specified with respect to the topology optimization problem of θ -type. It becomes apparent that the variation of θ which can be obtained from such methods have the regularity such that it could be included in the admissible set of design variable. In Sect. 8.7, the algorithm for solving the topology optimization problem of θ -type will be considered. However, the basic construction is the algorithm as shown in Sect. 3.7. Error estimations when numerical analyses are conducted via this algorithm are considered in Sect. 8.8. Here, the results of error estimation from the numerical analysis shown in Sect. 6.6 are used. Once the method for solving a Poisson-problem-related topology optimization problem of θ -type has been confirmed, we define a mean compliance minimization problem of a linear elastic body and an energy loss minimization problem of a Stokes flow field as topology optimization problems of θ -type and show the process for seeking the θ -derivatives of their cost functions in Sects. 8.9 and 8.10. Moreover, a numerical example with respect to a simple problem is shown in each section.

8.1 Set of Design Variables

Firstly, let us define a set of design variables in order to construct a topology optimization problem of a continuum. In this chapter, as shown in Fig. 8.3a, D is taken to be a $d \in \{2, 3\}$ -dimensional Lipschitz domain. $\Gamma_D \subset \partial D$ is taken to be a Dirichlet boundary and $|\Gamma_D| \neq 0$. $\Gamma_N \subset \partial D \setminus \bar{\Gamma}_D$ is a Neumann boundary.

In research so far, the range of density ϕ is limited to $[0, 1]$. The set of functions with restricted range such as this cannot be a linear space. Hence in this book, $\theta : D \rightarrow \mathbb{R}$ is set to be the design variable and the density is assumed to be given by a sigmoid function $\phi \in C^\infty(\mathbb{R}; \mathbb{R})$ with respect to θ . Several functions are known to

be sigmoid functions. Here, either

$$\phi(\theta) = \frac{1}{\pi} \tan^{-1} \theta + \frac{1}{2} \quad (8.1.1)$$

or

$$\phi(\theta) = \frac{1}{2} \tanh \theta + \frac{1}{2} \quad (8.1.2)$$

is used. These graphs are shown in Fig. 8.6. At this point, there is a need to note that $\phi : \mathbb{R} \rightarrow \mathbb{R}$ is a function which returns $(0, 1)$ when θ is given, and is not a function defined in D . However, because θ is a function with the domain D , $\phi(\theta)$ becomes a function defined in D .

With respect to this type of design variable θ , let us use the framework of an abstract optimal design problem (Problem 7.3.1) to define the linear space of design variables. As seen in Sect. 7.1, if the use of the gradient method is considered, the linear space of design variables needs to be a real Hilbert space. Hence, the linear space of design variable θ is set to be

$$X = \left\{ \theta \in H^1(D; \mathbb{R}) \mid \theta = 0 \text{ in } \bar{\Omega}_C \right\}, \quad (8.1.3)$$

where $\bar{\Omega}_C \subset \bar{D}$ is a boundary or a domain on which a variation of θ is compressed according to the design demands. If a function $\theta_C : D \rightarrow \mathbb{R}$ is specified and it is assumed that $\theta = \theta_C$ is established on $\bar{\Omega}_C$, $\tilde{\theta} = \theta - \theta_C$ is assumed to be an element of X . In particular, if $\bar{\Omega}_C$ is not required, $\theta \in X = H^1(D; \mathbb{R})$ is assumed.

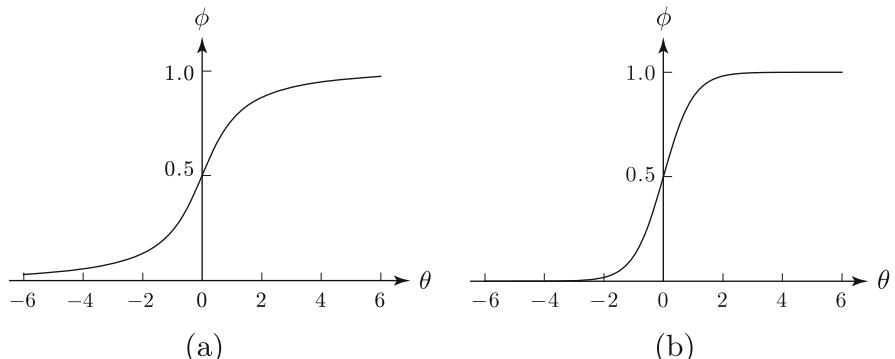


Fig. 8.6 Sigmoid functions for density $\phi(\theta)$ with respect to the design variable θ . (a) $\tan^{-1} \theta / \pi + 1/2$. (b) $(\tanh \theta + 1) / 2$

Furthermore, in order to be able to determine a Lipschitz continuous boundary from the isosurfaces of θ and to be compact in X , we assume that the admissible set of design variables is, at least, given by

$$\mathcal{D} = \left\{ \theta \in X \cap H^2(D; \mathbb{R}) \cap C^{0,1}(D; \mathbb{R}) \mid \max \{ \|\theta\|_{H^2(D; \mathbb{R})}, \|\theta\|_{C^{0,1}(D; \mathbb{R})} \} \leq \beta \right\} \quad (8.1.4)$$

where β is a positive constant. The requirement that \mathcal{D} is a compact set in X is assured by $H^2(D; \mathbb{R}) \Subset H^1(D; \mathbb{R})$ obtained from the Rellich–Kondrachov compact embedding theorem (Theorem 4.4.15). Moreover, in the same manner as Chap. 1, we consider that the boundedness constraint with norm is a side constraint and assume that θ is an interior point of \mathcal{D} ($\theta \in \mathcal{D}^\circ$) and when the side constraint is activated, we include it in the inequality constraints.

8.2 State Determination Problem

The linear space X and the admissible set \mathcal{D} of design variables have been defined, hence let us next define the boundary value problem of a partial differential equation which is a state determination problem. Here, the Poisson problem is considered for simplicity.

A Poisson problem (Problem 5.1.1) is defined in Chap. 5. Here, the Poisson problem when θ is a design variable is called the θ -type Poisson problem and its definition is shown based on the framework of an abstract optimal design problem (Problem 7.3.1).

The linear space of state variables (real Hilbert space) containing the homogeneous solution (which is given by $\tilde{u} = u - u_D$ with a known function u_D provided for Dirichlet condition) of θ -type Poisson problem is set to be

$$U = \left\{ u \in H^1(D; \mathbb{R}) \mid u = 0 \text{ on } \Gamma_D \right\}. \quad (8.2.1)$$

Furthermore, in order for the variation of θ obtained by the gradient method which will be shown later to be included in \mathcal{D} of Eq. (8.1.4), the admissible set of state variables \tilde{u} of the homogeneous form with respect to the state determination problem is set to be

$$\mathcal{S} = \left\{ u \in U \cap W^{1,2q_R}(D; \mathbb{R}) \mid \partial_\nu u|_{\Gamma_D} \in L^2(\Gamma_D; \mathbb{R}) \right\}, \quad (8.2.2)$$

where we let q_R be an integer satisfying $q_R > d$.

In order for the homogeneous solution \tilde{u} with respect to a state determination problem to be in \mathcal{S} , from the results seen in Sect. 5.3, the following assumptions are set with respect to the regularity of the known function.

Hypothesis 8.2.1 (Regularity of Given Functions) With respect to $q_R > d$, it is assumed that

$$b \in C^1 \left(X; L^{2q_R} (D; \mathbb{R}) \right), \quad p_N \in W^{1,2q_R} (D; \mathbb{R}), \quad u_D \in H^2 (D; \mathbb{R}),$$

where $C^1 (\cdot; \cdot)$ denotes the set of Fréchet differentiables (Definition 4.5.4). \square

Moreover, the following assumption is established with respect to the regularity of the boundary.

Hypothesis 8.2.2 (Opening Angle of Corner Point) Let D be a two-dimensional domain. In relation to the corner points on the boundary, and with respect to the Dirichlet boundary and Neumann boundary,

- (1) if the opening angle β is on the same type of boundary, $\beta < 2\pi$,
- (2) if it is on a mixed boundary, $\beta < \pi$.

Meanwhile, if D is a three-dimensional domain, the corner line on the boundary is smooth and it is assumed that the aforementioned relationship holds on the corner points of the boundary at a plane perpendicular to the corner line. Furthermore, the crossing points of the corner lines or the apexes of the conical boundaries are assumed not to have such singularities as they go beyond the framework within this book. \square

If Hypotheses 8.2.1 and 8.2.2 hold, the fact that u is included in $W^{1,2q_R} (D; \mathbb{R})$ can be confirmed as below. If Hypothesis 8.2.1 holds with respect to given functions, as seen in Sect. 5.3.1, u is included in $W^{1,2q_R}$ class at all points except those around the corner. Furthermore, from Proposition 5.3.1, if

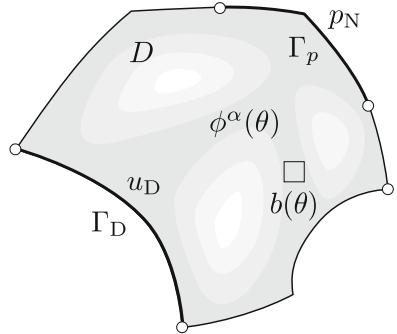
$$\omega > 1 - \frac{2}{2q_R} \quad (8.2.3)$$

holds, u becomes included in the $W^{1,2q_R}$ class. Here, in the neighborhood around corner points on the homogeneous boundary of Dirichlet or Neumann boundary type, $\omega = \pi/\beta$, hence the condition (1) in Hypothesis 8.2.2 can be obtained. Moreover, at the neighborhood of a corner point of the mixed boundary, from the fact that $\omega = \pi/(2\beta)$, the condition (2) of Hypothesis 8.2.2 is obtained.

Let us use the hypotheses above to define the state determination problem. For simplicity, we consider a Poisson problem such as in Fig. 8.7. Here, $\mathbf{v} \cdot \nabla$ is also to be written as ∂_v .

Problem 8.2.3 (θ -Type Poisson Problem) With respect to $\theta \in \mathcal{D}$, suppose that Hypotheses 8.2.1 and 8.2.2 are satisfied. Let $\alpha > 1$ be a constant and $\phi(\theta)$ a function given by Eq. (8.1.1) or Eq. (8.1.2). In this case, obtain $u : D \rightarrow \mathbb{R}$ which

Fig. 8.7 θ -type Poisson problem



satisfies

$$-\nabla \cdot (\phi^\alpha(\theta) \nabla u) = b(\theta) \quad \text{in } D,$$

$$\phi^\alpha(\theta) \partial_\nu u = p_N \quad \text{on } \Gamma_N,$$

$$u = u_D \quad \text{on } \Gamma_D. \quad \square$$

The unique existence of a weak solution to Problem 8.2.3 is guaranteed by the Lax–Milgram theorem (Theorem 5.2.4) with respect to $\tilde{u} = u - u_D \in U$. Subsequently, \tilde{u} will be used to mean $u - u_D$. Moreover, if Hypotheses 8.2.1 and 8.2.2 are satisfied, $u - u_D \in \mathcal{S}$ is guaranteed.

For later use, the Lagrange function with respect to Problem 8.2.3 is defined as

$$\begin{aligned} \mathcal{L}_S(\theta, u, v) &= \int_D (-\phi^\alpha(\theta) \nabla u \cdot \nabla v + b(\theta) v) \, dx \\ &+ \int_{\Gamma_N} p_N v \, d\gamma + \int_{\Gamma_D} \{(u - u_D) \phi^\alpha(\theta) \partial_\nu v + v \phi^\alpha(\theta) \partial_\nu u\} \, d\gamma, \end{aligned} \quad (8.2.4)$$

where u is not necessarily the solution of Problem 8.2.3, and $v \in \mathcal{S}$ was introduced as a Lagrange multiplier. That $v \in U$ is a Lagrange multiplier with respect to Problem 8.2.3 can be confirmed by recalling that in the process to obtain the weak form of the Poisson problem (Problem 5.1.1) in Chap. 5, $v \in U$ was introduced as a Lagrange multiplier. In Eq. (8.2.4), the third term on the right-hand side is a term which was removed in Chap. 5 using $u - u_D$, $v \in U$ when seeking the weak form of the Poisson problem. In reality, by removing this term, u and v could be viewed as H^1 -class functions (in order for $\partial_\nu v$ to have meaning on Γ_D , v needs to be a H^2 -class function). Here, however, that term will be left. The reason for this is because when a cost function f_i includes the boundary integral on Γ_D , it becomes apparent that the boundary condition of the adjoint problem with respect to f_i is seen from the boundary integral on Γ_D in the Lagrange function of f_i by using this

term. Matching the definition by Eq. (7.2.3) of a Lagrange function with respect to the abstract variational problem in Chap. 7, using $\tilde{u} = u - u_D$, we write

$$\mathcal{L}_S(\theta, u, v) = -a(\theta)(u, v) + l(\theta)(v) = -a(\theta)(\tilde{u}, v) + \hat{l}(\theta)(v), \quad (8.2.5)$$

where

$$a(\theta)(u, v) = \int_D \phi^\alpha(\theta) \nabla u \cdot \nabla v \, dx, \quad (8.2.6)$$

$$l(\theta)(v) = \int_D b(\theta) v \, dx + \int_{\Gamma_N} p_N v \, d\gamma, \quad (8.2.7)$$

$$\hat{l}(\theta)(v) = l(\theta)(v) + a(\theta)(u_D, v). \quad (8.2.8)$$

When u is the solution to Problem 8.2.3,

$$\mathcal{L}_S(\theta, u, v) = 0,$$

holds for all $v \in U$. This equation is equivalent to the weak form of Problem 8.2.3.

8.3 Topology Optimization Problem of θ -Type

The design variable θ and solution u of the state determination problem (state variable) were already defined. Hence, let us use these to define the topology optimization problem of θ -type. Here, we will consider a general cost function. Let u be the solution of a state determination problem (Problem 8.2.3) with respect to $\theta \in \mathcal{D}$ and set the cost function for each $i \in \{0, 1, \dots, m\}$ as

$$\begin{aligned} f_i(\theta, u) = & \int_D \zeta_i(\theta, u, \nabla u) \, dx + \int_{\Gamma_N} \eta_{Ni}(u) \, d\gamma \\ & - \int_{\Gamma_D} \eta_{Di}(\phi^\alpha(\theta) \partial_\nu u) \, d\gamma - c_i, \end{aligned} \quad (8.3.1)$$

where c_i is a constant and is determined so that some $(\theta, \tilde{u}) \in \mathcal{D} \times \mathcal{S}$ exists that satisfies $f_i \leq 0$ for all $i \in \{1, \dots, m\}$ (Slater constraint qualification is satisfied). Moreover, ζ_i , η_{Ni} and η_{Di} are assumed to be given by the following. These hypotheses will be used in an adjoint problem (Problem 8.5.1) to satisfy appropriate regularity requirements. To obtain the second-order θ derivatives of cost functions, additional hypotheses are needed but we will not specify them further. Nevertheless, we will only assume that sufficient conditions are satisfied by the state and adjoint state variables for us to be able to carry out a second-order differentiation of the cost functions with respect to θ .

Hypothesis 8.3.1 (Regularity of Cost Functions) For cost functions f_i ($i \in \{0, 1, \dots, m\}$) of Eq. (8.3.1), assume that $\zeta_i \in C^1(\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d; \mathbb{R})$, $\eta_{Ni} \in C^1(\mathbb{R}; \mathbb{R})$ and $\eta_{Di} \in C^1(\mathbb{R}; \mathbb{R})$, and with respect to $(\theta, u, \nabla u, \partial_\nu u) \in \mathcal{D} \times \mathcal{S} \times \mathcal{G} \times \mathcal{G}_{\Gamma_D}$ ($\mathcal{G} = \{\nabla u \mid u \in \mathcal{S}\}$, $\mathcal{G}_{\Gamma_D} = \{\partial_\nu u|_{\Gamma_D} \mid u \in \mathcal{S}\}$),

$$\begin{aligned} \zeta_i(\theta, u, \nabla u) &\in L^1(D; \mathbb{R}), \quad \zeta_{i\theta}(\theta, u, \nabla u) \in L^{q_R}(D; \mathbb{R}), \\ \zeta_{iu}(\theta, u, \nabla u) &\in L^{2q_R}(D; \mathbb{R}), \quad \zeta_{i(\nabla u)^\top}(\theta, u, \nabla u) \in W^{1,2q_R}(D; \mathbb{R}^d), \\ \eta_{Ni}(u) &\in L^1(\Gamma_N; \mathbb{R}), \quad \eta'_{Ni}(u) \in L^2(\Gamma_N; \mathbb{R}), \\ \eta_{Di}(\phi^\alpha(\theta) \partial_\nu u) &\in L^1(\Gamma_D; \mathbb{R}), \quad \eta'_{Di}(\phi^\alpha(\theta) \partial_\nu u) \in W^{1,2q_R}(\Gamma_D; \mathbb{R}). \quad \square \end{aligned}$$

As a supplementary explanation, if we are worried that $\partial_\nu u$ in the third term on the right-hand side of Eq. (8.3.1) cannot be defined from the assumption of $u - u_D \in \mathcal{S}$, it is remarked that this term will disappear due to the Dirichlet condition of adjoint problem (Problem 8.5.1) with respect to f_i shown later.

Using cost functions f_0, f_1, \dots, f_m of Eq. (8.3.1) and the framework of an abstract optimal design problem (Problem 7.3.1), the topology optimization problem of θ -type is defined as follows.

Problem 8.3.2 (Topology Optimization Problem of θ -Type) Let \mathcal{D} and \mathcal{S} be given by Eqs. (8.1.4) and (8.2.2), respectively. Suppose $f_0, \dots, f_m : X \times U \rightarrow \mathbb{R}$ are given by Eq. (8.3.1). In this case, obtain θ which satisfies

$$\min_{(\theta, u - u_D) \in \mathcal{D} \times \mathcal{S}} \{f_0(\theta, u) \mid f_1(\theta, u) \leq 0, \dots, f_m(\theta, u) \leq 0, \text{ Problem 8.2.3}\}. \quad \square$$

Based on the definition of a Lagrange function in Eq. (7.3.2) with respect to the abstract optimal design problem in Chap. 7, the Lagrange function with respect to Problem 8.3.2 is set to be

$$\begin{aligned} \mathcal{L}(\theta, u, v_0, v_1, \dots, v_m, \lambda_1, \dots, \lambda_m) \\ = \mathcal{L}_0(\theta, u, v_0) + \sum_{i \in \{1, \dots, m\}} \lambda_i \mathcal{L}_i(\theta, u, v_i), \end{aligned} \quad (8.3.2)$$

where $\lambda = \{\lambda_1, \dots, \lambda_m\}^\top \in \mathbb{R}^m$ are Lagrange multipliers with respect to $f_1 \leq 0, \dots, f_m \leq 0$. Moreover,

$$\begin{aligned} \mathcal{L}_i(\theta, u, v_i) &= f_i(\theta, u) + \mathcal{L}_S(\theta, u, v_i) \\ &= \int_D (\zeta_i(\theta, u, \nabla u) - \phi^\alpha(\theta) \nabla u \cdot \nabla v_i + b(\theta) v_i) \, dx \\ &\quad + \int_{\Gamma_N} (\eta_{Ni}(u) + p_N v_i) \, d\gamma \end{aligned}$$

$$\begin{aligned}
& + \int_{\Gamma_D} \left\{ (u - u_D) \phi^\alpha(\theta) \partial_\nu v_i \right. \\
& \left. + (v_i \phi^\alpha(\theta) \partial_\nu u - \eta_{D_i} (\phi^\alpha(\theta) \partial_\nu u)) \right\} d\gamma - c_i
\end{aligned} \quad (8.3.3)$$

is a Lagrange function with respect to f_i . Here, \mathcal{L}_S is the Lagrange function with respect to Problem 8.2.3 defined in Eq. (8.2.4). Moreover, v_i is a Lagrange multiplier with respect to a state determination problem prepared for f_i and assumes $v_i - \eta'_{D_i} \in U$. Furthermore, when thinking about the solution of the topology optimization problem of θ -type, the admissible set of $\tilde{v}_i = v_i - \eta'_{D_i}$ (admissible set of adjoint variables) needs to be a subset of \mathcal{S} .

8.4 Existence of an Optimum Solution

The existence of an optimum solution of Problem 8.3.2 can be assured by Theorem 7.4.4 in Chap. 7. To use it, we need to show the compactness of

$$\mathcal{F} = \{(\theta, \tilde{u}(\theta)) \in \mathcal{D} \times \mathcal{S} \mid \text{Problem 8.2.3}\} \quad (8.4.1)$$

and the (lower semi) continuity of f_0 . Here, we use $\tilde{u} = u - u_D \in U$.

The compactness of \mathcal{F} is presented in the following lemma.

Lemma 8.4.1 (Compactness of \mathcal{F}) *Suppose that Hypothesis 8.2.1 and Hypothesis 8.2.2 are satisfied. With respect to an arbitrary Cauchy sequence $\theta_n \rightarrow \theta$ in X which is uniformly convergent in \mathcal{D} and the solutions $\tilde{u}_n = \tilde{u}(\theta_n) \in U$ ($n \rightarrow \infty$) of Problem 8.2.3, the convergence*

$$\tilde{u}_n \rightarrow \tilde{u} \quad \text{strongly in } U$$

holds, and $\tilde{u} = \tilde{u}(\theta) \in U$ solves Problem 8.2.3. □

Proof Concerning the solution \tilde{u}_n of Problem 8.2.3 for θ_n , the inequality

$$\alpha_n \|\tilde{u}_n\|_U^2 \leq a(\theta_n)(\tilde{u}_n, \tilde{u}_n) = \hat{l}(\theta_n)(\tilde{u}_n) \leq \|\hat{l}(\theta_n)\|_{U'} \|\tilde{u}_n\|_U$$

holds, where $a(\theta_n)$ and $\hat{l}(\theta_n)$ are defined in Eq. (8.2.5), and α_n is a positive constant used in the definition of coerciveness for $a(\theta_n)$ (see (1) in the answer to Exercise 5.2.5). When $\theta_n \rightarrow \theta$ (uniform convergence in \mathcal{D}), α_n can be replaced by a positive constant α not depending with n . The norm $\|\hat{l}(\theta_n)\|_{U'} = \|l(\theta_n) + a(\theta_n)(u_D, \cdot)\|_{U'}$ ($l(\theta_n)$ is defined in Eq. (8.2.5)) being bounded can be shown using (3) in the answer to Exercise 5.2.5 via the convergence $\phi^\alpha(\theta_n) \rightarrow \phi^\alpha(\theta)$ for $\theta_n \rightarrow \theta$ (uniform convergence in \mathcal{D}), where $\hat{l}(v)$ and Ω in Exercise 5.2.5

are replaced by $\hat{l}(\theta_n)(v)$ and D , respectively. Hence, there exists a subsequence such that $\tilde{u}_n \rightarrow \tilde{u}$ weakly in U .

Next, we will show that \tilde{u} solves Problem 8.2.3 for θ . From the definition of Problem 8.2.3, we have

$$\lim_{n \rightarrow \infty} a(\theta_n)(\tilde{u}_n, v) = \lim_{n \rightarrow \infty} \hat{l}(\theta_n)(v), \quad (8.4.2)$$

with respect to an arbitrary $v \in U$. The right-hand side of Eq. (8.4.2) becomes

$$\lim_{n \rightarrow \infty} \hat{l}(\theta_n)(v) = \hat{l}(\theta)(v). \quad (8.4.3)$$

Indeed, from Hypothesis 8.2.1, the inequality

$$\begin{aligned} \left| \hat{l}(\theta_n)(v) - \hat{l}(\theta)(v) \right| &= \left| \int_D (b(\theta_n) - b(\theta)) v \, dx \right| \\ &\leq \|b(\theta_n) - b(\theta)\|_{L^2(D; \mathbb{R})} \|v\|_{L^2(D; \mathbb{R})} \rightarrow 0 \quad (n \rightarrow \infty) \end{aligned}$$

holds. The left-hand side of Eq. (8.4.2) becomes

$$\lim_{n \rightarrow \infty} a(\theta_n)(\tilde{u}_n, v) = a(\theta)(\tilde{u}, v). \quad (8.4.4)$$

This is due to the fact that, since $\tilde{u}_n \rightarrow \tilde{u}$ weakly in U , we then have

$$\begin{aligned} &|a(\theta_n)(\tilde{u}_n, v) - a(\theta)(\tilde{u}, v)| \\ &= \left| \int_D (\phi^\alpha(\theta_n) - \phi^\alpha(\theta)) \nabla \tilde{u}_n \cdot \nabla v \, dx \right| + \left| \int_D \phi^\alpha(\theta) \nabla(\tilde{u}_n - \tilde{u}) \cdot \nabla v \, dx \right| \\ &\leq \|\phi^\alpha(\theta_n) - \phi^\alpha(\theta)\|_{C^{0,1}(D; \mathbb{R})} \|\tilde{u}_n\|_{H^1(D; \mathbb{R})} \|v\|_{H^1(D; \mathbb{R})} + |a(\theta)(\tilde{u}_n - \tilde{u}, v)| \\ &\rightarrow 0 \quad (n \rightarrow \infty). \end{aligned}$$

Substituting Eqs. (8.4.3) and (8.4.4) into Eq. (8.4.2), the weak form of Problem 8.2.3 is obtained. It means that $\tilde{u} = \tilde{u}(\theta) \in U$ solves Problem 8.2.3.

Since the weak convergence of $\{\tilde{u}_n\}_{n \in \mathbb{N}}$ to \tilde{u} was shown, the strong convergence can be confirmed by showing

$$\|\tilde{u}_n\|_U \rightarrow \|\tilde{u}\|_U \quad (n \rightarrow \infty). \quad (8.4.5)$$

Indeed, when we use $a(\theta)$ in Eq. (8.2.6) as a norm in U and define

$$\|v\| = a(\theta)(v, v),$$

we have

$$\begin{aligned}
\|u_n\| &= a(\theta)(u_n, u_n) = a(\theta - \theta_n)(u_n, u_n) + a(\theta_n)(u_n, u_n) \\
&= \int_D (\phi^\alpha(\theta) - \phi^\alpha(\theta_n)) \nabla u_n \cdot \nabla u_n \, dx + l(\theta_n)(u_n) \\
&\rightarrow l(\theta)(u) = \|u\| \quad (n \rightarrow \infty).
\end{aligned} \tag{8.4.6}$$

From the above relation, it follows that $u_n \rightarrow u$ strongly in U , as desired. \square

We consider that the condition of $\tilde{u}(\theta)$ included in \mathcal{S} is guaranteed in the setting of Problem 8.2.3 (Hypotheses 8.2.1 and 8.2.2).

The continuity of f_0 means that f_0 is continuous on

$$S = \{(\theta, \tilde{u}(\theta)) \in \mathcal{F} \mid f_1(\theta, u(\theta)) \leq 0, \dots, f_m(\theta, u(\theta)) \leq 0\}. \tag{8.4.7}$$

Then, we will confirm the continuity of f_0 by showing the continuity of f_i ($i \in \{0, 1, \dots, m\}$) by the following lemma and assuming that S is not empty.

Lemma 8.4.2 (Continuity of f_0) *Let f_i be defined as in Eq. (8.3.1) under Hypothesis 8.3.1. Also, let $u_n \rightarrow u$ strongly in U which is determined by Lemma 8.4.1 with respect to an arbitrary Cauchy sequence $\theta_n \rightarrow \theta$ in X , which is uniformly convergent in \mathcal{D} , satisfy $\|\partial_v u_n - \partial_v u\|_{L^2(\Gamma_D; \mathbb{R})} \rightarrow 0$ ($n \rightarrow \infty$) on Γ_D . Then, f_i is continuous with respect to $\theta \in \mathcal{D}$.* \square

Proof The proof will be completed when

$$\begin{aligned}
|f_i(\theta_n, u_n) - f_i(\theta, u)| &\leq \left| \int_D (\zeta_i(\theta_n, u_n, \nabla u_n) - \zeta_i(\theta, u, \nabla u)) \, dx \right| \\
&\quad + \left| \int_{\Gamma_N} (\eta_{Ni}(u_n) - \eta_{Ni}(u)) \, d\gamma \right| \\
&\quad + \left| \int_{\Gamma_D} (\eta_{Di}(\phi^\alpha(\theta_n) \partial_v u_n) - \eta_{Di}(\phi^\alpha(\theta) \partial_v u)) \, d\gamma \right| \\
&= e_D + e_{\Gamma_N} + e_{\Gamma_D} \rightarrow 0 \quad (n \rightarrow \infty)
\end{aligned} \tag{8.4.8}$$

is shown with respect to $\theta_n \rightarrow \theta$ ($\theta_n, \theta \in \mathcal{D}$). From $\tilde{u}_n \rightarrow \tilde{u}$ weakly in U , the convergence $e_D \rightarrow 0$ ($n \rightarrow \infty$) is obtained. Indeed, writing $\tilde{\zeta}_i(t) = \zeta_i(t\theta_n + (1-t)\theta, tu_n + (1-t)u, t\nabla u_n + (1-t)\nabla u)$ ($t \in [0, 1]$) and considering $\zeta_i \in C^1(\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d; \mathbb{R})$ in Hypothesis 8.3.1, we get

$$\begin{aligned}
e_D &\leq \sup_{t \in [0, 1]} \left| \int_D \tilde{\zeta}_{i\theta}(t) [\theta_n - \theta] \, dx \right| + \sup_{t \in [0, 1]} \left| \int_D \tilde{\zeta}_{iu}(t) [u_n - u] \, dx \right| \\
&\quad + \sup_{t \in [0, 1]} \left| \int_D \tilde{\zeta}_i \nabla u(t) [\nabla u_n - \nabla u] \, dx \right|
\end{aligned}$$

$$\begin{aligned}
&\leq \sup_{t \in [0,1]} \left\| \tilde{\xi}_{i\theta}(t) \right\|_{L^{q_R}(D; \mathbb{R})} \|\theta_n - \theta\|_X + \sup_{t \in [0,1]} \left\| \tilde{\xi}_{iu}(t) \right\|_{L^{2q_R}(D; \mathbb{R})} \|u_n - u\|_U \\
&\quad + \sup_{t \in [0,1]} \left\| \tilde{\xi}_{i\nabla u}(t) \right\|_{W^{1,2q_R}(D; \mathbb{R}^d)} \|\nabla u_n - \nabla u\|_{L^2(D; \mathbb{R}^d)},
\end{aligned}$$

In the same way, we can obtain $e_{\Gamma_N} \rightarrow 0$. Meanwhile, the convergence $e_{\Gamma_D} \rightarrow 0$ ($n \rightarrow \infty$) can be shown in the following way. Writing $\tilde{\eta}_{Di}(t) = \eta_{Di}(\phi^\alpha(t\theta_n + (1-t)\theta)(t\partial_v u_n + (1-t)\partial_v u))$ ($t \in [0,1]$) and considering $\eta_{Di} \in C^1(\mathbb{R}; \mathbb{R})$ in Hypothesis 8.3.1, we see that the sequence of inequalities

$$\begin{aligned}
e_{\Gamma_D} &\leq \sup_{t \in [0,1]} \left| \int_{\Gamma_D} \tilde{\eta}'_{Di}(t) [\phi^\alpha(\theta_n) \partial_v u_n - \phi^\alpha(\theta) \partial_v u] dx \right| \\
&\leq \sup_{t \in [0,1]} \left| \int_{\Gamma_D} \tilde{\eta}'_{Di}(t) [(\phi^\alpha(\theta_n) - \phi^\alpha(\theta)) \partial_v u_n + \phi^\alpha(\theta) (\partial_v u_n - \partial_v u)] dx \right| \\
&\leq \sup_{t \in [0,1]} \left\| \tilde{\eta}'_{Di}(t) \right\|_{W^{1,2q_R}(D; \mathbb{R})} \\
&\quad \times (\|\theta_n - \theta\|_X \|\partial_v u\|_{L^2(\Gamma_D; \mathbb{R})} + \|\theta_n\|_X \|\partial_v u_n - \partial_v u\|_{L^2(\Gamma_D; \mathbb{R})})
\end{aligned}$$

holds. In conclusion, we obtain Eq. (8.4.8). \square

From Lemma 8.4.1, the compactness of \mathcal{F} was confirmed. The continuity of f_0 was shown by Lemma 8.4.2 and by the assumption that S is not empty. Then, under these conditions, it can be assured that there exists an optimum solution to Problem 8.3.2 by Theorem 7.4.4 (existence of an optimum solution).

Regarding the solution of Problem 8.3.2, we state the following remark.

Remark 8.4.3 (Existence of an Optimum Solution) The compactness of \mathcal{F} defined in Eq. (8.4.1) is based on the compactness of θ 's admissible set \mathcal{D} defined in Eq. (8.1.4). In Eq. (8.1.4), the condition $\max\{\|\theta\|_{H^2(D; \mathbb{R})}, \|\theta\|_{C^{0,1}(D; \mathbb{R})}\} \leq \beta$ with a positive constant β is added. This condition corresponds to the condition called a side constraint in Chap. 1. Such a side constraint is usually neglected, but it should be considered as a non-equality constraint when the condition becomes active. \square

Moreover, regarding the selection of the linear space X and the admissible set \mathcal{D} for the design variable, the following remark is left for reference.

Remark 8.4.4 (Selection of X and \mathcal{D}) The existence of a solution to Problem 8.3.2 was confirmed by Theorem 7.4.4 in Chap. 7. In the assumption of the theorem, it is necessary to assume that \mathcal{D} is a compact subset in X . In this chapter, this relation was satisfied by taking X as a function space of class H^1 and \mathcal{D} as a set of functions of $(H^2 \cap C^{0,1})$ -class based on the Rellich–Kondrachov compact embedding theorem (Theorem 4.4.15). However, there are other selections. For instance, it is possible to select X as a function space of C^0 class and \mathcal{D} as a set of functions of $C^{0,1}$ class. In this case, the Ascoli–Arzelà theorem (Theorem A.10.1) is used to show that \mathcal{D} is

a compact subset in X [62, proof in Theorem 2.1, p. 16]. When those are selected, the assumptions and lemmas are changed to show the existence of a solution. In this book, since a gradient method in a Hilbert space is considered, X and \mathcal{D} were selected as noted above. \square

8.5 Derivatives of Cost Functions

From this point onward, we will consider a solution to Problem 8.3.2 given that the conditions for its existence are satisfied. The Fréchet derivative of cost function f_i with respect to the variation of design variable θ will be referred to as a θ -derivative. Let us seek the θ -derivative of f_i by the Lagrange multiplier method such as that looked at in Sect. 7.5.2. Furthermore, let us seek the second-order θ -derivative of f_i using a method such as that seen in Sect. 7.5.3.

8.5.1 θ -Derivatives of Cost Functions

Let us focus on the Lagrange function \mathcal{L}_i of f_i defined in Eq. (8.3.3). The Fréchet derivative of \mathcal{L}_i with respect to an arbitrary variation $(\vartheta, \hat{u}, \hat{v}_i) \in X \times U \times U$ of (θ, u, v_i) becomes

$$\begin{aligned} \mathcal{L}'_i(\theta, u, v_i) & [\vartheta, \hat{u}, \hat{v}_i] \\ &= \mathcal{L}_{i\theta}(\theta, u, v_i) [\vartheta] + \mathcal{L}_{iu}(\theta, u, v_i) [\hat{u}] + \mathcal{L}_{iv_i}(\theta, u, v_i) [\hat{v}_i]. \end{aligned} \quad (8.5.1)$$

The third term on the right-hand side of Eq. (8.5.1) becomes

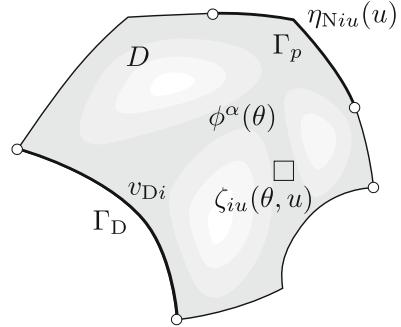
$$\mathcal{L}_{iv_i}(\theta, u, v_i) [\hat{v}_i] = \mathcal{L}_{Sv_i}(\theta, u, v_i) [\hat{v}_i] = \mathcal{L}_S(\theta, u, \hat{v}_i). \quad (8.5.2)$$

Equation (8.5.2) is the Lagrange function of the state determination problem (Problem 8.2.3). Here, if u is the weak solution of the state determination problem, the third term on the right-hand side of Eq. (8.5.1) is zero.

Moreover, the second term on the right-hand side of Eq. (8.5.1) becomes

$$\begin{aligned} \mathcal{L}_{iu}(\theta, u, v_i) & [\hat{u}] \\ &= \int_D \left(-\phi^\alpha(\theta) \nabla v_i \cdot \nabla \hat{u} + \zeta_{iu} \hat{u} + \zeta_{i(\nabla u)^\top} \cdot \nabla \hat{u} \right) dx \\ &+ \int_{\Gamma_N} \eta'_{Ni} \hat{u} d\gamma + \int_{\Gamma_D} \{ \hat{u} \phi^\alpha(\theta) \partial_\nu v + (v_i - \eta'_{Di}) \phi^\alpha(\theta) \partial_\nu \hat{u} \} d\gamma, \end{aligned} \quad (8.5.3)$$

Fig. 8.8 Adjoint problem with respect to f_i



where $\zeta_{iu}(\theta, u, \nabla u)[\hat{u}]$, $\zeta_{i(\nabla u)^\top}(\theta, u, \nabla u)[\nabla \hat{u}]$, $\eta'_{Ni}(u)[\hat{u}]$ and $\eta'_{Di}(u)[\phi^\alpha(\theta) \partial_v \hat{u}]$ were written as $\zeta_{iu} \hat{u}$, $\zeta_{i(\nabla u)^\top} \cdot \nabla \hat{u}$, $\eta'_{Ni} \hat{u}$ and $\eta'_{Di} \phi^\alpha(\theta) \partial_v \hat{u}$, respectively. Here, if v_i is determined so that Eq. (8.5.3) becomes zero, the second term on the right-hand side of Eq. (8.5.1) becomes zero. This relationship is the weak form of the adjoint problem with respect to f_i shown next. Here, when v_i is the weak solution of Problem 8.5.1, the second term of the right-hand side of Eq. (8.5.1) vanishes. The boundary condition of Problem 8.5.1 is as shown in Fig. 8.8.

Problem 8.5.1 (Adjoint Problem with Respect to f_i) For $\theta \in \mathcal{D}^\circ$, supposing the solution u is given to Problem 8.2.3, obtain $v_i : D \rightarrow \mathbb{R}$ which satisfies

$$\begin{aligned} -\nabla \cdot (\phi^\alpha(\theta) \nabla v_i) &= \zeta_{iu}(\theta, u, \nabla u) - \nabla \cdot (\zeta_{i(\nabla u)^\top}(\theta, u, \nabla u)) \quad \text{in } D, \\ \phi^\alpha(\theta) \partial_v v_i &= \eta'_{Ni}(u) + \zeta_{i(\nabla u)^\top} \cdot \mathbf{v} \quad \text{on } \Gamma_N, \\ v_i &= \eta'_{Di} \quad \text{on } \Gamma_D. \end{aligned}$$

□

Similarly to the solution u of the state determination problem, when Hypotheses 8.3.1 and 8.2.2 are satisfied, the solution $\tilde{v}_i = v_i - \eta'_{Di}$ of Problem 8.5.1 is guaranteed to be included in \mathcal{S} .

Furthermore, the first term on the right-hand side of Eq. (8.5.1) becomes

$$\begin{aligned} \mathcal{L}_{i\theta}(\theta, u, v_i)[\vartheta] &= \int_D \left(\zeta_{i\theta} + b' v_i - \alpha \phi^{\alpha-1} \phi' \nabla u \cdot \nabla v_i \right) \vartheta \, dx \\ &+ \int_{\Gamma_D} \alpha \phi^{\alpha-1} \phi' \left\{ (u - u_D) \partial_v v_i + (v_i - \eta'_{Di}) \partial_v u \right\} \, d\gamma. \end{aligned} \quad (8.5.4)$$

In the above equation, u and v_i are assumed to be the weak solutions of Problems 8.2.3 and 8.5.1, respectively. If we denote $f_i(\theta, u)$ here by $\tilde{f}_i(\theta)$, we can write

$$\tilde{f}'_i(\theta)[\vartheta] = \mathcal{L}_{i\theta}(\theta, u, v_i)[\vartheta] = \langle g_i, \vartheta \rangle, \quad (8.5.5)$$

where

$$g_i = \zeta_{i\theta} + b'v_i - \alpha\phi^{\alpha-1}\phi'\nabla u \cdot \nabla v_i. \quad (8.5.6)$$

When Eq. (8.1.1) is used in $\phi(\theta)$, we get

$$\phi'(\theta) = \frac{1}{\pi} \frac{1}{1+\theta^2}. \quad (8.5.7)$$

Moreover, when Eq. (8.1.2) is used,

$$\phi'(\theta) = \frac{1}{2} \operatorname{sech}^2 \theta = \frac{1}{2} \frac{1}{\cosh^2 \theta} = \frac{1}{(e^\theta + e^{-\theta})^2}. \quad (8.5.8)$$

From the above, the following results can be obtained with respect to θ -derivative g_i of f_i .

Theorem 8.5.2 (θ -Derivative of f_i) *For $\theta \in \mathcal{D}^\circ$, suppose u and v_i are the weak solutions of Problems 8.2.3 and 8.5.1 and these are said to be in \mathcal{S} of Eq. (8.2.2) (Hypotheses 8.2.1, 8.2.2 and 8.3.1 are satisfied). In this case, the θ -derivative of f_i becomes Eq. (8.5.5). Hence, g_i of Eq. (8.5.6) is in X' . Furthermore, $g_i \in L^{q_R}(D; \mathbb{R})$. \square*

Proof The fact that the θ -derivative of f_i becomes the g_i of Eq. (8.5.5) is as seen above. The following results can be obtained in respect of the regularity of g_i . If Hölder's inequality (Theorem A.9.1) and Poincaré's inequality (Corollary A.9.4) are used in Eq. (8.5.5),

$$\begin{aligned} & |\langle g_i, \vartheta \rangle|_{L^1(D; \mathbb{R})} \\ & \leq \left(\|\zeta_{i\theta}\|_{L^{q_R}(D; \mathbb{R})} + \|b'\|_{L^{2q_R}(D; \mathbb{R})} \|v_i\|_{L^{2q_R}(D; \mathbb{R})} \right. \\ & \quad \left. + \|\alpha\phi^{\alpha-1}\phi'\|_{C^\infty(\mathbb{R}; \mathbb{R})} \|\nabla u\|_{L^{2q_R}(D; \mathbb{R}^d)} \|\nabla v_i\|_{L^{2q_R}(D; \mathbb{R}^d)} \right) \|\vartheta\|_{L^2(D; \mathbb{R})} \\ & \leq \left(\|\zeta_{i\theta}\|_{L^{q_R}(D; \mathbb{R})} + \|b'\|_{L^{2q_R}(D; \mathbb{R})} \|v_i\|_{L^{2q_R}(D; \mathbb{R})} \right. \\ & \quad \left. + \|\alpha\phi^{\alpha-1}\phi'\|_{C^\infty(\mathbb{R}; \mathbb{R})} \|u\|_{W^{1,2q_R}(D; \mathbb{R})} \|v_i\|_{W^{1,2q_R}(D; \mathbb{R})} \right) \|\vartheta\|_X \end{aligned}$$

is obtained. The term (\cdot) on the right-hand side of the equation above is completely bounded due to the hypotheses. Hence, g_i is included in X' . Moreover, from the fact that each term within (\cdot) is in $L^{q_R}(D; \mathbb{R})$, $g_i \in L^{q_R}(D; \mathbb{R})$ can be obtained. \square

From Theorem 8.5.2, the following can be said about the regularity of the topology optimization problem of θ -type.

Remark 8.5.3 (Irregularity of Topology Optimization Problem of θ -Type) If in Theorem 8.5.2, Hypotheses 8.2.1, 8.2.2 and 8.3.1 are made more strict and a problem is constructed such that u and v_i are included in $W^{2,\infty}(D; \mathbb{R})$, g_i would be included in $C^{0,1}(D; \mathbb{R})$. In this case, the design variable $\theta + \epsilon \vartheta$ (ϵ is a positive constant) updated via the gradient method such that $-g_i$ is replaced by ϑ would be included in the admissible set of design variables \mathcal{D} . However, in such a case, it is necessary that the corner points permitted by Hypothesis 8.2.2 are removed, there is no boundary between the Dirichlet and Neumann boundaries, such as a Robin problem, or the neighborhoods of such points are included in $\bar{\Omega}_C$ in order to fix θ .

If these conditions are not satisfied, g_i is not included in $C^{0,1}(D; \mathbb{R})$. Hence, in a gradient method such that $-g_i$ is replaced by ϑ , $\theta + \epsilon \vartheta$ is not included in the admissible set \mathcal{D} of design variables. This result is thought to be one of the reasons for numerical instability phenomena in which a checkerboard pattern appears such as that in Fig. 8.5. \square

8.5.2 Second-Order θ -Derivative of Cost Functions

Furthermore, let us seek the second-order derivative (Hessian) of the cost function with respect to the variation of the design variable. In Sect. 7.5.3, the way to seek a second-order Fréchet derivative with respect to an abstract optimal design problem has already been shown. Hence, let us follow that method in order to seek the second-order θ -derivative of \tilde{f}_i with respect to f_i given in Eq. (8.3.1).

The following assumption is established in order to obtain the second-order θ -derivative of \tilde{f}_i .

Hypothesis 8.5.4 (Second-Order θ -Derivative of \tilde{f}_i) With respect to the state determination problem (Problem 8.2.3) and the cost function f_i defined in Eq. (8.3.1), assume respectively that:

- (1) b is not a function of θ ,
- (2) ζ_i is not a function of u (it is a function of θ and ∇u). \square

Hypothesis 8.5.4 will be used in Eq. (8.5.16) to obtain Eq. (8.5.17). However, in the method shown in Sect. 8.5.3, this hypothesis will not be required.

The Lagrange function \mathcal{L}_i of f_i is defined by Eq. (8.3.3). Viewing (θ, u) as a design variable and putting its admissible set and admissible set of directions as

$$S = \{(\theta, u) \in \mathcal{D} \times \mathcal{S} \mid \mathcal{L}_S(\theta, u, v) = 0 \text{ for all } v \in U\},$$

$$T_S(\theta, u) = \{(\vartheta, \hat{v}) \in X \times U \mid \mathcal{L}_{S\theta u}(\theta, u, v)[\vartheta, \hat{v}] = 0 \text{ for all } v \in U\},$$

the second-order Fréchet partial derivative of \mathcal{L}_i with respect to arbitrary variations $(\vartheta_1, \hat{v}_1), (\vartheta_2, \hat{v}_2) \in T_S(\theta, u)$ of $(\theta, u) \in S$, similarly to Eq. (7.5.21), becomes

$$\begin{aligned} & \mathcal{L}_{i(\theta, u)(\theta, u)}(\theta, u, v_i)[(\vartheta_1, \hat{v}_1), (\vartheta_2, \hat{v}_2)] \\ &= \mathcal{L}_{i\theta\theta}(\theta, u, v_i)[\vartheta_1, \vartheta_2] + \mathcal{L}_{i\theta u}(\theta, u, v_i)[\vartheta_1, \hat{v}_2] \\ & \quad + \mathcal{L}_{i\theta u}(\theta, u, v_i)[\vartheta_2, \hat{v}_1] + \mathcal{L}_{i u u}(\theta, u, v_i)[\hat{v}_1, \hat{v}_2]. \end{aligned} \quad (8.5.9)$$

Each term on the right-hand side of Eq. (8.5.9) becomes

$$\mathcal{L}_{i\theta\theta}(\theta, u, v_i)[\vartheta_1, \vartheta_2] = \int_D \left\{ \zeta_{i\theta\theta} - (\phi^\alpha(\theta))'' \nabla u \cdot \nabla v_i \right\} \vartheta_1 \vartheta_2 \, dx, \quad (8.5.10)$$

$$\begin{aligned} & \mathcal{L}_{i\theta u}(\theta, u, v_i)[\vartheta_1, \hat{v}_2] \\ &= \int_D \left\{ \zeta_{i\theta(\nabla u)^\top} \cdot \nabla \hat{v}_2 - (\phi^\alpha(\theta))' \nabla \hat{v}_2 \cdot \nabla v_i \right\} \vartheta_1 \, dx, \end{aligned} \quad (8.5.11)$$

$$\begin{aligned} & \mathcal{L}_{i u \theta}(\theta, u, v_i)[\vartheta_2, \hat{v}_1] \\ &= \int_D \left\{ \zeta_{i\theta(\nabla u)^\top} \cdot \nabla \hat{v}_1 - (\phi^\alpha(\theta))' \nabla \hat{v}_1 \cdot \nabla v_i \right\} \vartheta_2 \, dx, \end{aligned} \quad (8.5.12)$$

$$\mathcal{L}_{i u u}(\theta, u, v_i)[\hat{v}_1, \hat{v}_2] = 0. \quad (8.5.13)$$

Here, the fact that $u = u_D$, $v_i = \eta'_{D i}$, \hat{v}_1 and \hat{v}_2 become zero on Γ_D was used. Moreover,

$$(\phi^\alpha(\theta))' = \alpha \phi^{\alpha-1}(\theta) \phi'(\theta), \quad (8.5.14)$$

$$(\phi^\alpha(\theta))'' = \alpha(\alpha-1) \phi^{\alpha-2}(\theta) \phi'^2(\theta) + \alpha \phi^{\alpha-1}(\theta) \phi''(\theta). \quad (8.5.15)$$

On the other hand, with respect to arbitrary variations $(\vartheta_j, \hat{v}_j) \in T_S(\theta, u)$ for $j \in \{1, 2\}$, the Fréchet partial derivative of \mathcal{L}_S becomes

$$\begin{aligned} & \mathcal{L}_{S\theta u}(\theta, u, v)[\vartheta_j, \hat{v}_j] \\ &= \int_D \left\{ -(\phi^\alpha(\theta))' \vartheta \nabla u - \phi^\alpha(\theta) \nabla \hat{v}_j \right\} \cdot \nabla v \, dx \\ &= 0 \end{aligned} \quad (8.5.16)$$

with respect to an arbitrary $v \in U$. Here, Hypothesis 8.5.4 and the fact that v and \hat{v}_j are both zero on Γ_D were used. From Eq. (8.5.16), we can obtain

$$\nabla \hat{v}_j = -\frac{(\phi^\alpha(\theta))'}{\phi^\alpha(\theta)} \vartheta_j \nabla u \quad \text{in } D. \quad (8.5.17)$$

This relation becomes possible the following argument.

Substituting \hat{v}_j of Eq. (8.5.17) into \hat{v}_j in Eq. (8.5.9), and considering Dirichlet boundary conditions in the state determination problem and the adjoint problems with respect to f_1, \dots, f_m , as well as $\hat{v}_j = 0$ on Γ_D , we have

$$\begin{aligned} \mathcal{L}_{i\theta u}(\theta, u, v_i)[\vartheta_1, \vartheta_2] &= \mathcal{L}_{iu\theta}(\theta, u, v_i)[\hat{v}_1, \vartheta_2] \\ &= \int_D \frac{(\phi^\alpha(\theta))'}{\phi^\alpha(\theta)} \left\{ (\phi^\alpha(\theta))' \nabla v_i - \zeta_{i\theta(\nabla u)^\top} \right\} \cdot \nabla u \vartheta_1 \vartheta_2 \, dx. \end{aligned} \quad (8.5.18)$$

Summarizing the results above, from Eqs. (8.5.10), (8.5.13) and (8.5.18), the second-order θ -derivative of \tilde{f}_i becomes

$$\begin{aligned} h_i(\theta, u, v_i)[\vartheta_1, \vartheta_2] &= \int_D \left[\left\{ 2 \frac{(\phi^\alpha(\theta))'^2}{\phi^\alpha(\theta)} - (\phi^\alpha(\theta))'' \right\} \nabla u \cdot \nabla v_i \right. \\ &\quad \left. + \zeta_{i\theta\theta} - 2 \frac{(\phi^\alpha(\theta))'}{\phi^\alpha(\theta)} \zeta_{i\theta(\nabla u)^\top} \cdot \nabla u \right] \vartheta_1 \vartheta_2 \, dx \\ &= \int_D \left(\beta(\alpha, \theta) \nabla u \cdot \nabla v_i + \zeta_{i\theta\theta} - 2\alpha \frac{\phi'(\theta)}{\phi(\theta)} \zeta_{i\theta(\nabla u)^\top} \cdot \nabla u \right) \vartheta_1 \vartheta_2 \, dx, \end{aligned} \quad (8.5.19)$$

where

$$\beta(\alpha, \theta) = \alpha(\alpha+1) \phi^{\alpha-2}(\theta) \phi'^2(\theta) - \alpha \phi^{\alpha-1}(\theta) \phi''(\theta). \quad (8.5.20)$$

When $\phi(\theta)$ is given by Eq. (8.1.1) or Eq. (8.1.2), they respectively become

$$\begin{aligned} \beta(\alpha, \theta) &= \alpha(\alpha+1) \left(\frac{1}{\pi} \tan^{-1} \theta + \frac{1}{2} \right)^{\alpha-2} \left\{ \frac{1}{\pi(1+\theta^2)} \right\}^2 \\ &\quad - \alpha \left(\frac{1}{\pi} \tan^{-1} \theta + \frac{1}{2} \right)^{\alpha-1} \left\{ -\frac{2\theta}{\pi(1+\theta^2)^2} \right\} \end{aligned} \quad (8.5.21)$$

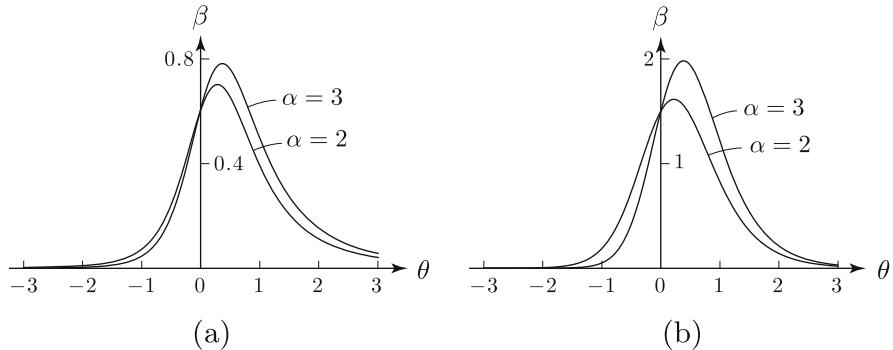


Fig. 8.9 Coefficient function $\beta(\alpha, \theta)$ in the second-order θ -derivative of the cost function. (a) $\phi(\theta) = \tan^{-1} \theta / \pi + 1/2$. (b) $\phi(\theta) = (\tanh \theta + 1) / 2$

or

$$\begin{aligned} \beta(\alpha, \theta) = \alpha(\alpha+1) & \left(\frac{1}{2} \tanh \theta + \frac{1}{2} \right)^{\alpha-2} \left(\frac{\operatorname{sech}^2 \theta}{2} \right)^2 \\ & - \alpha \left(\frac{1}{2} \tanh \theta + \frac{1}{2} \right)^{\alpha-1} (-\operatorname{sech}^2 \theta \tanh \theta). \end{aligned} \quad (8.5.22)$$

Figure 8.9 shows the graph of $\beta(\alpha, \theta)$. From these graphs, the fact that $\beta(\alpha, \theta) > 0$ holds can be confirmed. Furthermore, if the remainder term in (·) on the right-hand side of Eq. (8.5.19) is positive and bounded, $h_i(\theta, u, v_i)[\cdot, \cdot]$ becomes a coercive and bounded bilinear form on X .

8.5.3 Second Order θ -Derivative of Cost Function Using Lagrange Multiplier Method

When the Lagrange multiplier method is used to obtain the second-order θ -derivative of a cost function, we use the same idea as proposed in Sect. 7.5.4. Fixing ϑ_1 , we define the Lagrange function with respect to $\tilde{f}_i'(\theta)[\vartheta_1] = \langle g_i, \vartheta_1 \rangle$ in Eq. (8.5.5) by

$$\mathcal{L}_{li}(\theta, u, v_i, w_i, z_i) = \langle g_i, \vartheta_1 \rangle + \mathcal{L}_S(\theta, u, w_i) + \mathcal{L}_{Ai}(\theta, v_i, z_i), \quad (8.5.23)$$

where \mathcal{L}_S is given by Eq. (8.2.4), and

$$\begin{aligned} \mathcal{L}_{Ai}(\theta, v_i, z_i) &= \int_D \left(-\phi^\alpha(\theta) \nabla v_i \cdot \nabla z_i + \zeta_{iu} z_i + \zeta_{i(\nabla u)^\top} \cdot \nabla z_i \right) dx \\ &\quad + \int_{\Gamma_N} \eta'_{Ni} z_i \, d\gamma + \int_{\Gamma_D} \{z_i \phi^\alpha(\theta) \partial_\nu v + (v_i - \eta'_{Di}) \phi^\alpha(\theta) \partial_\nu z_i\} \, d\gamma \end{aligned} \quad (8.5.24)$$

is the Lagrange function with respect to the adjoint problem (Problem 8.5.1) with respect to f_i . $w_i \in U$ and $z_i \in U$ are the adjoint variables provided for u and v_i in g_i .

With respect to arbitrary variations $(\vartheta_2, \hat{u}, \hat{v}_i, \hat{w}_i, \hat{z}_i) \in X \times U^4$ of $(\theta, u, v_i, w_i, z_i)$, the Fréchet derivative of \mathcal{L}_{li} is written as

$$\begin{aligned} \mathcal{L}'_{li}(\theta, u, v_i, w_i, z_i) &[\vartheta_2, \hat{u}, \hat{v}_i, \hat{w}_i, \hat{z}_i] \\ &= \mathcal{L}_{li\theta}(\theta, u, v_i, w_i, z_i) [\vartheta_2] + \mathcal{L}_{liu}(\theta, u, v_i, w_i, z_i) [\hat{u}] \\ &\quad + \mathcal{L}_{liv_i}(\theta, u, v_i, w_i, z_i) [\hat{v}_i] + \mathcal{L}_{liw_i}(\theta, u, v_i, w_i, z_i) [\hat{w}_i] \\ &\quad + \mathcal{L}_{liz_i}(\theta, u, v_i, w_i, z_i) [\hat{z}_i]. \end{aligned} \quad (8.5.25)$$

The fourth term on the right-hand side of Eq. (8.5.25) vanishes if u is the solution of the state determination problem. If v_i can be determined as the solution of the adjoint problem, the fifth term of Eq. (8.5.25) also vanishes.

The second term on the right-hand side of Eq. (8.5.25) is

$$\begin{aligned} \mathcal{L}_{liu}(\theta, u, v_i, w_i, z_i) &[\hat{u}] \\ &= \int_D \left[\left\{ \left(\zeta_{i\theta u} + \zeta_{iuu} z_i + \zeta_{i(\nabla u)^\top u} \cdot \nabla z_i \right) \hat{u} - \alpha \phi^{\alpha-1} \phi' \nabla v_i \cdot \nabla \hat{u} \right\} \vartheta_1 \right. \\ &\quad \left. - \phi^\alpha \nabla w_i \cdot \nabla \hat{u} \right] dx. \end{aligned} \quad (8.5.26)$$

Here, the condition that Eq. (8.5.26) is zero for arbitrary $\hat{u} \in U$ is equivalent to setting w_i to be the solution of the following adjoint problem.

Problem 8.5.5 (Adjoint Problem of w_i with Respect to (g_i, ϑ_1)) Under the assumption of Problem 8.3.2, letting $\vartheta_1 \in X$ be given, find $w_i = w_i(\vartheta_1) \in U$ satisfying

$$\begin{aligned} -\nabla \cdot (\phi^\alpha \nabla w_i) &= \left(\nabla \cdot \left(\alpha \phi^{\alpha-1} \phi' \nabla v_i \right) + \zeta_{i\theta u} + \zeta_{iuu} z_i + \zeta_{i(\nabla u)^\top u} \cdot \nabla z_i \right) \vartheta_1 \\ &\quad \text{in } D, \end{aligned}$$

$$\phi^\alpha \partial_\nu w_i = \alpha \phi^{\alpha-1} \phi' \partial_\nu v_i \vartheta_1 \quad \text{on } \Gamma_N,$$

$$w_i = 0 \quad \text{on } \Gamma_D. \quad \square$$

The third term on the right-hand side of Eq. (8.5.25) is

$$\begin{aligned} & \mathcal{L}_{li v_i}(\theta, u, v_i, w_i, z_i) [\hat{v}_i] \\ &= \int_D \left\{ \left(b' \hat{v}_i - \alpha \phi^{\alpha-1} \phi' \nabla u \cdot \nabla \hat{v}_i \right) \vartheta_1 - \phi^\alpha \nabla z_i \cdot \nabla \hat{v}_i \right\} dx. \end{aligned} \quad (8.5.27)$$

Here, the condition that Eq. (8.5.27) is zero for arbitrary $\hat{v}_i \in U$ is equivalent to setting z_i to be the solution of the following adjoint problem.

Problem 8.5.6 (Adjoint Problem of z_i with Respect to (g_i, ϑ_1)) Under the assumption of Problem 8.3.2, letting $\vartheta_1 \in X$ be given, find $z_i = z_i(\vartheta_1) \in U$ satisfying

$$\begin{aligned} -\nabla \cdot (\phi^\alpha \nabla z_i) &= \left(\nabla \cdot \left(\alpha \phi^{\alpha-1} \phi' \nabla u \right) + b' \right) \vartheta_1 \quad \text{in } D, \\ \phi^\alpha \partial_v z_i &= \alpha \phi^{\alpha-1} \phi' \partial_v u \vartheta_1 \quad \text{on } \Gamma_N, \\ z_i &= 0 \quad \text{on } \Gamma_D. \end{aligned}$$

□

Finally, the first term on the right-hand side of Eq. (8.5.25) becomes

$$\begin{aligned} & \mathcal{L}_{li \theta}(\theta, u, v_i, w_i, z_i) [\vartheta_2] \\ &= \int_D \left[\left\{ \zeta_{i \theta \theta} u + b'' v_i - \left(\alpha(\alpha-1) \phi^{\alpha-2} \phi'^2 + \alpha \phi^{\alpha-1} \phi'' \right) \nabla u \cdot \nabla v_i \right\} \vartheta_1 \right. \\ & \quad \left. - \alpha \phi^{\alpha-1} \phi' (\nabla u \cdot \nabla w_i + \nabla v_i \cdot \nabla z_i) + b' (w_i + z_i) \right] \vartheta_2 dx. \end{aligned}$$

Here, u, v_i, w_i (ϑ_1) and z_i (ϑ_1) are assumed to be the weak solutions of Problems 8.2.3, 8.5.1, 8.5.5 and 8.5.6, respectively. If we denote $f_i(\theta, u)$ here by $\tilde{f}_i(\theta)$, we have the relation:

$$\begin{aligned} \mathcal{L}_{li \theta}(\theta, u, v_i, w_i(\vartheta_1), z_i(\vartheta_1)) [\vartheta_2] &= \tilde{f}_i''(\theta) [\vartheta_1, \vartheta_2] \\ &= \langle g_{Hi}(\theta, \vartheta_1), \vartheta_2 \rangle, \end{aligned} \quad (8.5.28)$$

where the Hesse gradient g_{Hi} of f_i is given by

$$\begin{aligned} g_{Hi}(\theta, \vartheta_1) &= \left\{ - \left(\alpha(\alpha-1) \phi^{\alpha-2} \phi'^2 + \alpha \phi^{\alpha-1} \phi'' \right) \nabla u \cdot \nabla v_i + \zeta_{i \theta \theta} u + b'' v_i \right\} \vartheta_1 \\ & \quad - \alpha \phi^{\alpha-1} \phi' (\nabla u \cdot \nabla w_i(\vartheta_1) + \nabla v_i \cdot \nabla z_i(\vartheta_1)) \\ & \quad + b' (w_i(\vartheta_1) + z_i(\vartheta_1)). \end{aligned} \quad (8.5.29)$$

We obtained two different expressions for the second-order derivative $\tilde{f}_i''(\theta)[\vartheta_1, \vartheta_2]$. Under the assumption of Hypothesis 8.5.4, they accord when using the same ϑ_1 and ϑ_2 . This relation will be confirmed in Sects. 8.9 and 8.10.

8.6 Descent Directions of Cost Functions

In Remark 8.5.3, it was shown that the topology optimization problem of θ -type becomes irregular unless a special assumption of regularity is set. Here, let us consider the gradient method and Newton method on the linear space X of design variables with the functionality of regularizing θ -derivative of a cost function. Here, with respect to the $i \in \{0, \dots, m\}$ th cost function f_i , assume that the gradient $g_i \in X'$ of Eq. (8.5.6) and Hessian $h_i \in \mathcal{L}^2(X \times X; \mathbb{R})$ of Eq. (8.5.19) are given and think about the method to obtain a descent direction of f_i using the gradient method and Newton method on the linear space X of design variables.

8.6.1 H^1 Gradient Method

The method for obtaining the aforementioned descent direction vector using the solution $\vartheta_{gi} \in X$ to the next problem will be referred to as the H^1 gradient method of θ -type.

Problem 8.6.1 (H^1 Gradient Method of θ -Type) Let X and \mathcal{D} be Eqs. (8.1.3) and (8.1.4), respectively. Let $a_X : X \times X \rightarrow \mathbb{R}$ be a bounded and coercive bilinear form on X . In other words, it is supposed that some positive constants α_X and β_X exist and that

$$a_X(\vartheta, \vartheta) \geq \alpha_X \|\vartheta\|_X^2, \quad |a_X(\vartheta, \psi)| \leq \beta_X \|\vartheta\|_X \|\psi\|_X \quad (8.6.1)$$

holds with respect to arbitrary $\vartheta \in X$ and $\psi \in X$. For each $f_i \in C^1(\mathcal{D}; \mathbb{R})$, let $g_i(\theta_k) \in X'$ be its corresponding θ -derivative at $\theta_k \in \mathcal{D}^\circ$ which is not a local minimum point. In this case, obtain $\vartheta_{gi} \in X$ which satisfies

$$a_X(\vartheta_{gi}, \psi) = -\langle g_i(\theta_k), \psi \rangle \quad (8.6.2)$$

with respect to an arbitrary $\psi \in X$. □

There is an arbitrary property for choosing $a_X : X \times X \rightarrow \mathbb{R}$ which was assumed in Problem 8.6.1. Several specific examples will be shown below.

Method Using the Inner Product of H^1 Space

An inner product in a real Hilbert space is coercive. Hence, let us use the inner product as

$$a_X(\vartheta, \psi) = \int_D (\nabla \vartheta \cdot \nabla \psi + c_D \vartheta \psi) \, dx, \quad (8.6.3)$$

Here, it is assumed that c_D is a uniformly bounded element of $L^\infty(D; \mathbb{R})$ which is positive almost everywhere. In this case a_X is a coercive bilinear form on X (see solution to Exercise 5.2.7 (1)). Moreover, the following can be said for the way to choose c_D . If c_D takes a large value, the second term in the integral of the right-hand side of Eq. (8.6.3) is dominant compared to the first term and suppresses the smoothing functionality. Hence, it becomes a result closer to when $-g_i$ is chosen to be the direct search vector. Here, the size of the search vector (step size) is considered to be adjusted by the size of the positive constant c_a used in the algorithm in Sect. 7.7.1 (same as Sect. 3.7).

The strong form of H^1 gradient method when using Eq. (8.6.3) is as below.

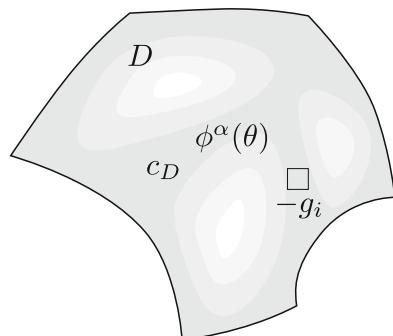
Problem 8.6.2 (H^1 Gradient Method Using H^1 Inner Product) Let $\theta \in \mathcal{D}^\circ$, and $g_i \in X'$ of Eq. (8.5.6) be given. Find $\vartheta_{gi} : D \rightarrow \mathbb{R}$ which satisfies

$$\begin{aligned} -\Delta \vartheta_{gi} + c_D \vartheta_{gi} &= -g_i && \text{in } D, \\ \partial_\nu \vartheta_{gi} &= 0 && \text{on } \partial D. \end{aligned}$$

□

Figure 8.10 shows the image of Problem 8.6.2. This problem is a boundary value problem of an elliptic partial differential equation when Ω in the extended Poisson problem of Problem 5.1.3 has changed to D , and $c_{\partial D} = 0$ is set. Hence, numerical solutions can be obtained via numerical analysis methods such as the finite element method.

Fig. 8.10 H^1 gradient method using inner product in H^1 space



Method Using Boundary Conditions

Moreover, even if the Dirichlet condition or Robin condition with respect to θ are used, the bilinear form $a_X : X \times X \rightarrow \mathbb{R}$ can be made coercive.

Firstly, let us think about using the Dirichlet boundary condition. On Eq. (8.1.3) where the linear space X of the design variable is defined, $\bar{\Omega}_C \subset \bar{D}$ was defined to be a boundary or domain in which θ is fixed under the design demand. Here, the measure of the boundary or domain for $\bar{\Omega}_C$ is assumed to have a positive value. In this case,

$$a_X(\vartheta, \psi) = \int_{D \setminus \bar{\Omega}_C} \nabla \vartheta \cdot \nabla \psi \, dx \quad (8.6.4)$$

is a bounded and coercive bilinear form on X as seen in the solution for Exercise 5.2.5. The strong form equation of H^1 gradient method in this case is as follows.

Problem 8.6.3 (H^1 Gradient Method Using Dirichlet Condition) Let $g_i \in X'$ of Eq. (8.5.6) be given with respect to $\theta \in \mathcal{D}^\circ$. Obtain $\vartheta_{g_i} : D \setminus \bar{\Omega}_C \rightarrow \mathbb{R}$ which satisfies

$$-\Delta \vartheta_{g_i} = -g_i \quad \text{in } D \setminus \bar{\Omega}_C,$$

$$\partial_\nu \vartheta_{g_i} = 0 \quad \text{on } \partial D \setminus \bar{\Omega}_C,$$

$$\vartheta_{g_i} = 0 \quad \text{in } \bar{\Omega}_C.$$

□

Problem 8.6.3 is a problem replacing Ω and Γ_D in the Poisson problem of Problem 5.1.1 with $D \setminus \bar{\Omega}_C$ and $\partial \Omega_C$, respectively. Figure 8.11a shows its image. The numerical solution of this problem can also be obtained via numerical analysis method such as the finite element method.

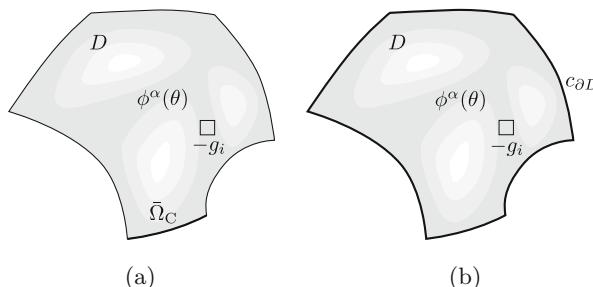


Fig. 8.11 H^1 gradient method using boundary conditions **(a)** Dirichlet condition. **(b)** Robin condition

Furthermore, if the Robin condition is used, even if $\bar{\Omega}_C = \emptyset$ is assumed in Eq. (8.1.3), the coerciveness of $a_X(\vartheta, \psi)$ can be obtained. Some positive-valued and uniformly bounded function $c_{\partial D} \in L^\infty(\partial D; \mathbb{R})$ is chosen and

$$a_X(\vartheta, \psi) = \int_D \nabla \vartheta \cdot \nabla \psi \, dx + \int_{\partial D} c_{\partial D} \vartheta \psi \, d\gamma. \quad (8.6.5)$$

The fact that this a_X becomes a coercive bilinear form in X is shown in the solution of Exercise 5.2.7 (2). In this case, the strong form is as follows.

Problem 8.6.4 (H^1 Gradient Method Using the Robin Condition) Let $g_i \in X'$ of Eq. (8.5.6) be given at $\theta \in \mathcal{D}^\circ$. Obtain $\vartheta_{gi} : D \rightarrow \mathbb{R}$ which satisfies

$$\begin{aligned} -\Delta \vartheta_{gi} &= -g_i && \text{in } D, \\ \partial_\nu \vartheta_{gi} + c_{\partial D} \vartheta_{gi} &= 0 && \text{on } \partial D. \end{aligned}$$

□

Figure 8.11b shows an image of Problem 8.6.4. The Robin condition for this problem is a condition typically used for a heat transfer boundary when a Poisson problem is viewed as a stationary heat transfer problem. The external temperature at the boundary is set to zero and the heat transfer coefficient is set to be $c_{\partial D}$. Here, the numerical solutions of this problem can be obtained via a numerical analysis method, such as the finite element method.

Regularity of H^1 Gradient Method

The following results can be obtained with respect to the weak solutions of the H^1 gradient method (Problems 8.6.2 to 8.6.4) with respect to the topology optimization problem of θ -type. Here, in this section, vicinities of the singular points as follows are denoted as B : when D is a two-dimensional domain, concave corner points on ∂D and corner points on $\partial\Gamma_D$ in mixed boundary conditions for which the opening angle is greater than $\pi/2$, and when D is a three-dimensional domain, concave edges on ∂D and edges on $\partial\Gamma_D$ in mixed boundary conditions for which the opening angle is greater than $\pi/2$. Moreover, $f_i(\theta, u)$ when u is the solution to Problem 8.2.3 is written as $\tilde{f}_i(\theta)$.

Theorem 8.6.5 (H^1 Gradient Method of θ -Type) *With respect to $g_i \in L^{q_R}(D; \mathbb{R})$ in Theorem 8.5.2, the weak solutions ϑ_{gi} of Problems 8.6.2 to 8.6.4 exist uniquely, and ϑ_{gi} is in the $H^2(D; \mathbb{R}) \cap C^{0,1}$ class on $D \setminus \bar{B}$. Moreover, ϑ_{gi} is a descent direction for this function.* □

Proof From the fact that g_i is in $L^{q_R}(D; \mathbb{R}) \subset X'$, the Lax–Milgram theorem says that the weak solutions ϑ_{gi} of Problems 8.6.2 to 8.6.4 uniquely exist. Moreover, the following results can be obtained regarding the regularity of the solution ϑ_{gi} . From the fact that ϑ_{gi} satisfies an elliptic partial differential equation, the differentiability increases by two orders compared to g_i ; it becomes $W^{2,q_R} \subset H^2(D; \mathbb{R})$ class on

$D \setminus \bar{B}$. If Sobolev's embedding theorem (Theorem 4.3.14) is applied to this, when $q_R > d$,

$$2 - \frac{d}{q_R} = 1 + \sigma > 1$$

holds, where $\sigma \in (0, 1)$. Therefore, in Theorem 4.3.14 (3), when $p = q_R$, $q = \infty$, $k = 1$ and $j = 1$,

$$W^{2,q_R}(D \setminus \bar{B}, \mathbb{R}) \subset C^{0,1}(D \setminus \bar{B}, \mathbb{R})$$

holds on $D \setminus \bar{B}$. Then, ϑ_{gi} becomes $H^2(D; \mathbb{R}) \cap C^{0,1}$ class on $D \setminus \bar{B}$. Furthermore, with respect to the weak solutions ϑ_{gi} of Problems 8.6.2 to 8.6.4 and for some positive constant $\bar{\epsilon}$, the estimate

$$\begin{aligned} \tilde{f}_i(\theta + \bar{\epsilon} \vartheta_{gi}) - \tilde{f}_i(\theta) &= \bar{\epsilon} \langle \mathbf{g}_i, \vartheta_{gi} \rangle + o(|\bar{\epsilon}|) = -\bar{\epsilon} \alpha_X(\vartheta_{gi}, \vartheta_{gi}) + o(|\bar{\epsilon}|) \\ &\leq -\bar{\epsilon} \alpha_X \|\vartheta_{gi}\|_X^2 + o(|\bar{\epsilon}|) \end{aligned}$$

holds. Here, if $\bar{\epsilon}$ is taken to be sufficiently small, $\tilde{f}_i(\theta)$ decreases. \square

If the direction of variation of the design variable is determined using the H^1 gradient method, a solution is found in the admissible set \mathcal{D} of design variables excepting the neighborhood of singular points from Theorem 8.6.5. From this, it is thought that the H^1 gradient method is a regular gradient method.

8.6.2 H^1 Newton Method

Furthermore, if it is possible to calculate the second-order derivative (Hessian) $h_i \in \mathcal{L}^2(X \times X; \mathbb{R})$ of the cost function f_i , the Newton method on $X = H^1(D; \mathbb{R})$ can be considered. Such a method is referred to as the H^1 Newton method of θ -type.

Problem 8.6.6 (H^1 Newton Method of θ -Type) Let X and \mathcal{D} be Eqs. (8.1.3) and (8.1.4), respectively. For $f_i \in C^2(\mathcal{D}; \mathbb{R})$, its θ -derivative and second-order θ -derivative at $\theta_k \in \mathcal{D}^\circ$, which is not a local minimum point, are taken respectively to be $g_i(\theta_k) \in X'$ and $h_i(\theta_k) \in \mathcal{L}^2(X \times X; \mathbb{R})$. Moreover, let $\alpha_X : X \times X \rightarrow \mathbb{R}$ be a coercive and bounded bilinear form on X . Here, obtain $\vartheta_{gi} \in X$ which satisfies

$$h_i(\theta_k)[\vartheta_{gi}, \psi] + \alpha_X(\vartheta_{gi}, \psi) = -\langle g_i(\theta_k), \psi \rangle \quad (8.6.6)$$

with respect to an arbitrary $\psi \in X$. \square

In Problem 8.6.6, if the left-hand side of Eq. (8.6.6) is made to be just h_i , it cannot be expected to fix the irregularity of $g_i(\theta_k)$ pointed out in Remark 8.5.3. In reality,

h_i calculated with Eq. (8.5.19) does not include the term of $\nabla \vartheta_1 \cdot \nabla \vartheta_2$. Hence, in Problem 8.6.6, a bilinear form a_X was added in order to ensure coerciveness and boundedness of the left-hand side of Eq. (8.6.6) on X and the regularity of ϑ_{gi} . For example, if Eq. (8.6.3) using an inner product on X is to be used as a basis, let:

$$a_X(\vartheta, \psi) = \int_D (c_{D1} \nabla \vartheta \cdot \nabla \psi + c_{D0} \vartheta \psi) \, dx. \quad (8.6.7)$$

Here, c_{D0} and c_{D1} are positive constants for achieving the coerciveness and regularity respectively. These have the same meaning as that explained after Eq. (8.6.3).

Furthermore, in the case of the Newton method when the second-order θ -derivative of $f_i(\theta)$ is given by the Hesse gradient, Problem 8.6.6 is replaced with the following problem.

Problem 8.6.7 (Newton Method of θ -Type Using Hesse Gradient) Let X and \mathcal{D} be Eqs. (8.1.3) and (8.1.4), respectively. For $f_i \in C^2(\mathcal{D}; \mathbb{R})$, the gradient of the θ -derivative of f_i , search vector, which is obtained in the previous step by the H^1 gradient method or H^1 Newton method of θ -type using the Hesse gradient, and Hesse gradient of f_i at a non-local minimum point $\theta_k \in \mathcal{D}^\circ$ are denoted as $g_i(\theta_k) \in X'$, $\bar{\vartheta}_{gi} \in X$ and $g_{Hi}(\theta_k, \bar{\vartheta}_{gi}) \in X'$, respectively. $a_X : X \times X \rightarrow \mathbb{R}$ is a coercive and bounded bilinear form on X . Here, obtain a $\vartheta_{gi} \in X$ which satisfies

$$a_X(\vartheta_{gi}, \psi) = - \langle (g_i(\theta_k) + g_{Hi}(\theta_k, \bar{\vartheta}_{gi})), \psi \rangle \quad (8.6.8)$$

with respect to an arbitrary $\psi \in X$. \square

8.7 Solution of Topology Optimization Problem of θ -Type

The abstract optimal design problem (Problem 7.3.1) and topology optimization problem of θ -type (Problem 8.3.2) can be dealt with as in Table 8.1. Therefore by appropriate replacements, the gradient method and Newton method with respect to constrained problems shown in Sect. 7.7.1 (Sect. 3.7) and Sect. 7.7.2 (Sect. 3.8) can be applied.

Table 8.1 Correspondence between abstract optimal design problem (Problem 7.3.1) and topology optimization problem of θ -type (Problem 8.3.2)

	Abstract problem	Topology optimization problem
Design variable	$\phi \in X$	$\theta \in X = H^1(D; \mathbb{R})$
State variable	$u \in U$	$u \in U = H^1(D; \mathbb{R})$
Fréchet derivative of f_i	$g_i \in X'$	$g_i \in X' = H^{1/2}(D; \mathbb{R})$
Solution of gradient method	$\varphi_{gi} \in X$	$\vartheta_{gi} \in X = H^1(D; \mathbb{R})$

8.7.1 Gradient Method for Constrained Problems

The gradient method with respect to a constrained problem can have a simple algorithm such as Algorithm 3.6 shown in Sect. 3.7.1, which can be used by applying changes such as those below:

- (1) Replace the design variable \mathbf{x} and its variation \mathbf{y} as θ and ϑ , respectively.
- (2) Equation (3.7.10) providing the gradient method is replaced by conditions that establish

$$c_a a_X (\vartheta_{g_i}, \psi) = -\langle g_i, \psi \rangle \quad (8.7.1)$$

with respect to an arbitrary $\psi \in X$, where $a_X (\vartheta_{g_i}, \psi)$ is a bilinear form on X used as the weak form of Problems 8.6.2 to 8.6.4.

- (3) Replace Eq. (3.7.11) seeking the search vector with

$$\vartheta_g = \vartheta_{g0} + \sum_{i \in I_A} \lambda_i \vartheta_{g_i}, \quad (8.7.2)$$

where $I_A = \left\{ i \in \{1, \dots, m\} \mid \tilde{f}_i(\theta) \geq 0 \right\}$.

- (4) Replace Eq. (3.7.12) seeking the Lagrange multipliers with

$$(\langle g_i, \vartheta_{gj} \rangle)_{(i,j) \in I_A^2} (\lambda_j)_{j \in I_A} = - (f_i + \langle g_i, \vartheta_{g0} \rangle)_{i \in I_A}. \quad (8.7.3)$$

Equation (8.7.3) is solved possibly several times, removing each time the constraints where the associated Lagrange multiplier is negative (active set method).

Furthermore, if a complicated algorithm such as Algorithm 3.7 is to be used, the following changes should be added to (1) to (4) above:

- (5) The Armijo criterion Eq. (3.7.26) is replaced, with respect to $\xi \in (0, 1)$, with

$$\mathcal{L}(\theta + \vartheta_g, \lambda_{k+1}) - \mathcal{L}(\theta, \lambda) \leq \xi \left\langle g_0 + \sum_{i \in I_A} \lambda_i g_i, \vartheta_g \right\rangle. \quad (8.7.4)$$

- (6) Replace the Wolfe criterion Eq. (3.7.27), with respect to μ ($0 < \xi < \mu < 1$), with

$$\begin{aligned} & \mu \left\langle g_0 + \sum_{i \in I_A} \lambda_i g_i, \vartheta_g \right\rangle \\ & \leq \left\langle g_0 (\theta + \vartheta_g) + \sum_{i \in I_A} \lambda_{i+1} g_i (\theta + \vartheta_g), \vartheta_g \right\rangle. \end{aligned} \quad (8.7.5)$$

- (7) Replace Eq. (3.7.21) for updating λ_{k+1} based on the Newton–Raphson method by

$$\begin{aligned}\delta\lambda &= (\delta\lambda_j)_{j \in I_A} \\ &= -\left(\langle g_i(\lambda_{k+1}), \vartheta_{gj}(\lambda_{k+1}) \rangle\right)^{-1}_{(i,j) \in I_A^2} (f_i(\lambda_{k+1}))_{i \in I_A}.\end{aligned}\quad (8.7.6)$$

Let us look at the points to be aware of when solving a topology optimization problem of θ -type such as the one above.

In the topology optimization problem of θ -type (Problem 8.3.2), the solution of the state determination problem with respect to a design variable $\theta \in X$ or cost function becomes a non-convex non-linear mapping. This is because the coefficient $\phi^\alpha(\theta)$ of a partial differential equation used in a SIMP model is a composite function of a sigmoid function and a power function. Hence, depending on the definitions of cost functions and boundary conditions, there may be cases when several local minimum points exist. In that case, the initial distribution of θ needs to be changed and the convergence results need to be compared.

Moreover, the initial distribution of θ must be in the admissible set \mathcal{D} of design variables defined in Eq. (8.1.4). In other words, it needs to be a continuous function. If the initial distribution of θ is given by the characteristic function (an L^∞ class function) corresponding to the location of some holes, caution needs to be taken that the discontinuities of θ at the boundaries of the holes are not removed, even when the methods above are used.

Furthermore, in topology optimization problems of θ -type shown in this chapter, a sigmoid function is used to change θ to ϕ . Therefore, as ϕ nears 0 and 1, there is a disadvantage that the gradient of ϕ with respect to θ becomes small and convergence is slowed. This issue will hopefully be improved using the Newton method shown in the next section.

8.7.2 Newton Method for Constrained Problems

If the second-order θ -derivatives are computable in addition to the θ -derivatives of the cost functions, the gradient method with respect to a constrained problem can be changed to a Newton method with respect to a constrained problem. In this case, $h_i(\theta_k)[\vartheta_{gi}, \psi]$ of Eq. (8.6.6) is replaced by

$$h_{\mathcal{L}}(\theta_k)[\vartheta_{gi}, \psi] = h_0(\theta_k)[\vartheta_{gi}, \psi] + \sum_{i \in I_A(\theta_k)} \lambda_{ik} h_i(\theta_k)[\vartheta_{gi}, \psi]. \quad (8.7.7)$$

In other words, let Eq. (8.6.6) be

$$c_h h_{\mathcal{L}}(\theta_k)[\vartheta_{gi}, \psi] + a_X(\vartheta_{gi}, \psi) = -\langle g_i(\theta_k), \psi \rangle, \quad (8.7.8)$$

where a_X is defined by Eq. (8.6.7), and c_h , c_{D0} and c_{D1} are positive constants to control the step size. Under this situation, a simple algorithm such as Algorithm 3.8 shown in Sect. 3.8.1 can be utilized after the following replacements:

- (1) Replace the design variable x and its variation y with θ and ϑ respectively.
- (2) Replace Eq. (3.7.10) with the solution of Eq. (8.7.8).
- (3) Replace Eq. (3.7.11) with Eq. (8.7.2).
- (4) Replace Eq. (3.7.12) with Eq. (8.7.3).

When the second-order θ -derivative of $f_i(\theta)$ is obtained as a Hesse gradient, Eqs. (8.7.7) and (8.7.8) are replaced with

$$g_{H\mathcal{L}}(\theta_k, \bar{\vartheta}_g) = g_{H0}(\theta_k, \bar{\vartheta}_g) + \sum_{i \in I_A(\theta_k)} \lambda_{ik} g_{Hi}(\theta_k, \bar{\vartheta}_g), \quad (8.7.9)$$

$$a_X(\vartheta_{gi}, \psi) = -\langle (g_i(\theta_k) + c_h g_{H\mathcal{L}}(\theta_k, \bar{\vartheta}_g)), \psi \rangle, \quad (8.7.10)$$

respectively. Using the definitions, the following is added:

- (5) Replace Eq. (3.8.11) with Eq. (8.7.10).

Furthermore, when considering the complicated algorithm shown in Sect. 3.8.2, all sorts of innovations in response to the additional functionalities and characteristics of problems become necessary.

8.8 Error Estimation

If an algorithm such as the one shown in Sect. 8.7 is to be used to solve the topology optimization problem of θ -type (Problem 8.3.2), the search vector ϑ_g can be obtained via Eq. (8.7.2). In this regard, there is a need to obtain the solutions to the boundary value problems of three elliptic partial differential equations. In other words, the solution u to the state determination problem (Problem 8.2.3), the solutions $v_0, v_{i_1}, \dots, v_{i_{|I_A|}}$ of the adjoint problems (Problem 8.5.1) with respect to $f_0, f_{i_1}, \dots, f_{i_{|I_A|}}$ and the solutions $\vartheta_0, \vartheta_{i_1}, \dots, \vartheta_{i_{|I_A|}}$ from the H^1 gradient method of θ -type (Problem 8.6.1). Furthermore, there is a need to seek the Lagrange multipliers $\lambda_{i_1}, \dots, \lambda_{i_{|I_A|}}$ with Eq. (8.7.3). Here, the numerical solutions with respect to the three boundary value problems are assumed to be obtained by the finite element method, so let us use the results of error estimation with respect to the numerical solutions from the finite element method looked at in Sect. 6.6 in order to conduct the error estimation of the search vector ϑ_g [120, 121].

Furthermore, if instead of the H^1 gradient method, the H^1 Newton method is to be used, evaluations of the second-order derivatives of cost functions are required. However, these will be omitted for now.

In this section, D is assumed to be a polygon in two dimensions; a polyhedron in three dimensions and a regular finite element division $\mathcal{T} = \{D_i\}_{i \in \mathcal{E}}$ with respect to

D is considered. Moreover, we define the maximum diameter h of finite elements as $h(\mathcal{T})$ of Eq. (6.6.2) and consider a sequence $\{\mathcal{T}_h\}_{h \rightarrow 0}$ of finite element divisions. Hereinafter, notation such as that below will be used:

- (1) Let the exact solutions of a state determination problem (Problem 8.2.3) and adjoint problems (Problem 8.5.1) with respect to $f_0, f_{i_1}, \dots, f_{i_{|I_A|}}$ be u and $v_0, v_{i_1}, \dots, v_{i_{|I_A|}}$, respectively. Moreover, their numerical solutions from the finite element method can be written, with respect to $i \in I_A \cup \{0\}$, as

$$u_h = u + \delta u_h, \quad (8.8.1)$$

$$v_{ih} = v_i + \delta v_{ih}. \quad (8.8.2)$$

- (2) Let the numerical solutions of θ -derivatives of cost functions $f_0, f_{i_1}, \dots, f_{i_{|I_A|}}$ be

$$g_{ih} = g_i + \delta g_{ih}, \quad (8.8.3)$$

where g_i and g_{ih} are functions of $u, v_0, v_{i_1}, \dots, v_{i_{|I_A|}}$ and $u_h, v_{0h}, v_{i_1h}, \dots, v_{i_{|I_A|}h}$, respectively.

- (3) Let the exact solutions from the H^1 gradient method (for example, Problem 8.6.2) calculated using the exact solutions $g_0, g_{i_1}, \dots, g_{i_{|I_A|}}$ of θ -derivatives be $\vartheta_{g0}, \vartheta_{gi_1}, \dots, \vartheta_{gi_{|I_A|}}$. Moreover, the exact solutions from the H^1 gradient method calculated using the numerical solutions $g_{0h}, g_{i_1h}, \dots, g_{i_{|I_A|}h}$ are written, with respect to $i \in I_A \cup \{0\}$, as

$$\hat{\vartheta}_{gi} = \vartheta_{gi} + \delta \hat{\vartheta}_{gi}. \quad (8.8.4)$$

- (4) Let the numerical solutions from the H^1 gradient method calculated using the numerical solutions $g_{0h}, g_{i_1h}, \dots, g_{i_{|I_A|}h}$, with respect to $i \in I_A \cup \{0\}$, be

$$\vartheta_{gih} = \hat{\vartheta}_{gi} + \delta \hat{\vartheta}_{gih} = \vartheta_{gi} + \delta \vartheta_{gih}. \quad (8.8.5)$$

- (5) The coefficient matrix $(\langle g_i, \vartheta_{gj} \rangle)_{(i,j) \in I_A^2}$ of Eq. (8.7.3) constructed using $g_0, g_{i_1}, \dots, g_{i_{|I_A|}}$ and $\vartheta_{g0}, \vartheta_{gi_1}, \dots, \vartheta_{gi_{|I_A|}}$ as \mathbf{A} . Moreover, the coefficient matrix $(\langle g_{ih}, \vartheta_{gjh} \rangle)_{(i,j) \in I_A^2}$ of Eq. (8.7.3) constructed from $g_{0h}, g_{i_1h}, \dots, g_{i_{|I_A|}h}$ and $\vartheta_{g0h}, \vartheta_{gi_1h}, \dots, \vartheta_{gi_{|I_A|}h}$ is denoted as $\mathbf{A}_h = \mathbf{A} + \delta \mathbf{A}_h$. Here, we assume that $f_i = 0$ and denote $-(\langle g_i, \vartheta_{g0} \rangle)_{i \in I_A}$ as \mathbf{b} . Moreover, $-(\langle g_{ih}, \vartheta_{g0h} \rangle)_{i \in I_A}$ is denoted as $\mathbf{b}_h = \mathbf{b} + \delta \mathbf{b}_h$. Furthermore, the exact solution of the Lagrange multiplier is written as $\boldsymbol{\lambda} = \mathbf{A}^{-1} \mathbf{b}$. Furthermore, its numerical solution is written as

$$\boldsymbol{\lambda}_h = (\lambda_{ih})_{i \in I_A} = \mathbf{A}_h^{-1} \mathbf{b}_h = \boldsymbol{\lambda} + \delta \boldsymbol{\lambda}_h. \quad (8.8.6)$$

- (6) Equation (8.7.2) constructed by $\lambda_{i_1 h}, \dots, \lambda_{i_{|I_A|} h}$ and numerical solutions $\vartheta_{g0h}, \vartheta_{gi_1h}, \dots, \vartheta_{gi_{|I_A|}h}$ is written as

$$\vartheta_{gh} = \vartheta_{g0h} + \sum_{i \in I_A} \lambda_{ih} \vartheta_{gih} = \vartheta_g + \delta \vartheta_{gh}. \quad (8.8.7)$$

In the definition above, $\delta \vartheta_{gh}$ defined in Eq. (8.8.7) represents the error of the search vector. In this section, the aim is to evaluate the order of its norm $\|\delta \vartheta_{gh}\|_X$ with respect to h . Here, the following hypothesis is established.

Hypothesis 8.8.1 (Error Estimation of ϑ_g) In a state determination problem and adjoint problems with respect to $f_0, f_{i_1}, \dots, f_{i_{|I_A|}}$, let $\alpha > 1$. Moreover, the following is assumed with respect to $q_R > d$ and $k, j \in \{1, 2, \dots\}$:

- (1) The homogeneous form for the exact solutions u of a state determination problem and $v_0, v_{i_1}, \dots, v_{i_{|I_A|}}$ of adjoint problems with respect to $f_0, f_{i_1}, \dots, f_{i_{|I_A|}}$ are elements of

$$\mathcal{S}_k = U \cap W^{k+1, 2q_R}(D; \mathbb{R}). \quad (8.8.8)$$

In order for this condition to be established, Hypotheses 8.2.1, 8.2.2 and 8.3.1 need to be amended.

- (2) The integrand of the cost function f_i is, with respect to $i \in I_A \cup \{0\}$,

$$\zeta_{i\theta u} \in L^{2q_R}(D; \mathbb{R}).$$

- (3) There exist positive constants c_1, c_2, c_3 which do not depend on h and for $i \in I_A \cup \{0\}$,

$$\|\delta u_h\|_{W^{j, 2q_R}(D; \mathbb{R})} \leq c_1 h^{k+1-j} |u|_{W^{k+1, 2q_R}(D; \mathbb{R})}, \quad (8.8.9)$$

$$\|\delta v_{ih}\|_{W^{j, 2q_R}(D; \mathbb{R})} \leq c_2 h^{k+1-j} |v_i|_{W^{k+1, 2q_R}(D; \mathbb{R})}, \quad (8.8.10)$$

$$\|\delta \hat{\vartheta}_{gih}\|_{W^{j, 2q_R}(D; \mathbb{R})} \leq c_3 h^{k+1-j} |\hat{\vartheta}_{gi}|_{W^{k+1, 2q_R}(D; \mathbb{R})}, \quad (8.8.11)$$

is satisfied, where $|\cdot|$ expresses a semi-norm (see Eq. (4.3.12)).

- (4) With respect to the coefficient matrix \mathbf{A}_h in Eq. (8.8.6), a positive constant c_4 exists and

$$\|\mathbf{A}_h^{-1}\|_{\mathbb{R}^{|I_A| \times |I_A|}} \leq c_4$$

is satisfied, where $\|\cdot\|_{\mathbb{R}^{|I_A| \times |I_A|}}$ represents the norm of a matrix (see Eq. (4.4.3)). \square

In (1) of Hypothesis 8.8.1, since $k \in \{1, 2, \dots\}$, it is a stronger condition than \mathcal{S} defined in Eq. (8.2.2). The reason for this is because in (3) of Hypothesis 8.8.1, u and

$v_0, v_{i_1}, \dots, v_{i_{|I_A|}}$ on the right-hand side of Eqs. (8.8.9) and (8.8.10) need to be in the $W^{k+1,2q_R}$ class. Hypothesis 8.8.1 (3) is based on Corollary 6.6.4. Hypothesis 8.8.1 (4) is a condition which is established when $g_{i_1}, \dots, g_{i_{|I_A|}}$ are linearly independent.

Here, Theorem 8.8.5 shown later can be obtained. In order to show this result, the following three lemmas are used.

Lemma 8.8.2 (Error Estimation of g_i) *When the assumptions (1) and (2) in Hypothesis 8.8.1 as well as Eqs. (8.8.9) and (8.8.10) are satisfied, with respect to δg_{ih} of Eq. (8.8.3), there exists a positive constant c_5 which does not depend on h and the estimate*

$$\langle \delta g_{ih}, \vartheta \rangle \leq c_5 h^k \|\vartheta\|_X$$

is established with respect to an arbitrary $\vartheta \in X$. \square

Proof δg_{ih} is yielded by the numerical error of δu_h and δv_{ih} . Hence, from Eq. (8.5.5),

$$|\langle \delta g_i, \vartheta \rangle| \leq |\mathcal{L}_{i\theta u}(\theta, u, v_i)[\vartheta, \delta u_h] + \mathcal{L}_{i\theta v_i}(\theta, u, v_i)[\vartheta, \delta v_{ih}]| \quad (8.8.12)$$

is established. If with respect to the right-hand side of Eq. (8.8.12), Hölder's inequality (Theorem A.9.1) and Poincaré's inequality (Corollary A.9.4) are used,

$$\begin{aligned} & |\mathcal{L}_{i\theta u}(\theta, u, v_i)[\vartheta, \delta u_h] + \mathcal{L}_{i\theta v_i}(\theta, u, v_i)[\vartheta, \delta v_{ih}]| \\ & \leq \left\{ \|\zeta_{i\theta u}\|_{L^{2q_R}(D; \mathbb{R})} \|\delta u_h\|_{L^{2q_R}(D; \mathbb{R})} + \|b'\|_{L^{2q_R}(D; \mathbb{R})} \|\delta v_{ih}\|_{L^{2q_R}(D; \mathbb{R})} \right. \\ & \quad + \|\alpha\phi^{\alpha-1}\phi'\|_{L^\infty(D; \mathbb{R})} \|\nabla \delta u_h\|_{L^{2q_R}(D; \mathbb{R}^d)} \|\nabla v_i\|_{L^{2q_R}(D; \mathbb{R}^d)} \\ & \quad \left. + \|\alpha\phi^{\alpha-1}\phi'\|_{L^\infty(D; \mathbb{R})} \|\nabla u_h\|_{L^{2q_R}(D; \mathbb{R}^d)} \|\nabla \delta v_{ih}\|_{L^{2q_R}(D; \mathbb{R}^d)} \right\} \|\vartheta\|_X \\ & \leq \left\{ \|\zeta_{i\theta u}\|_{L^{2q_R}(D; \mathbb{R})} \|\delta u_h\|_{W^{1,2q_R}(D; \mathbb{R})} + \|b'\|_{L^{2q_R}(D; \mathbb{R})} \|\delta v_{ih}\|_{W^{1,2q_R}(D; \mathbb{R})} \right. \\ & \quad + \|\alpha\phi^{\alpha-1}\phi'\|_{L^\infty(D; \mathbb{R})} \|\delta u_h\|_{W^{1,2q_R}(D; \mathbb{R})} \|v_i\|_{W^{1,2q_R}(D; \mathbb{R})} \\ & \quad \left. + \|\alpha\phi^{\alpha-1}\phi'\|_{L^\infty(D; \mathbb{R})} \|u_h\|_{W^{1,2q_R}(D; \mathbb{R})} \|\delta v_{ih}\|_{W^{1,2q_R}(D; \mathbb{R})} \right\} \|\vartheta\|_X \end{aligned}$$

is established. Here, Eqs. (8.8.9) and (8.8.10) in which $j = 1$ as well as Hypothesis 8.8.1 (1) leads to the result of the lemma. \square

Lemma 8.8.3 (Error Estimation of ϑ_{gi}) *When Hypothesis 8.8.1 (1), (2) and (3) are satisfied, with respect to $\delta\vartheta_{gi}$ of Eq. (8.8.5), the positive constant c_6 which does not depend on h exists and the inequality*

$$\|\delta\vartheta_{gih}\|_X \leq c_6 h^k$$

is established. □

Proof The following is established from Eqs. (8.8.4) and (8.8.5):

$$\|\delta\vartheta_{gih}\|_X \leq \|\delta\hat{\vartheta}_{gi}\|_X + \|\delta\hat{\vartheta}_{gih}\|_X. \quad (8.8.13)$$

Here $\|\delta\hat{\vartheta}_{gi}\|_X$ represents the error of the exact solution of the H^1 gradient method (for example, Problem 8.6.2) by δg_{ih} in Lemma 8.8.2 and $\|\delta\hat{\vartheta}_{gih}\|_X$ represents the error with respect to the numerical solution of H^1 gradient method. $\|\delta\hat{\vartheta}_{gih}\|_X$ of Eq. (8.8.13) satisfies

$$a_X(\delta\hat{\vartheta}_{gi}, \vartheta) = -\langle \delta g_{ih}, \vartheta \rangle$$

with respect to an arbitrary $\vartheta \in X$. Here, if $\vartheta = \delta\hat{\vartheta}_{gi}$, the bound

$$\alpha_X \|\delta\hat{\vartheta}_{gi}\|_X^2 \leq |\langle \delta g_{ih}, \delta\hat{\vartheta}_{gi} \rangle| \quad (8.8.14)$$

is established, where α_X is a positive constant used in Eq. (8.6.1). If Lemma 8.8.2 is used with respect to δg_{ih} of Eq. (8.8.14), the estimate

$$\|\delta\hat{\vartheta}_{gi}\|_X \leq \frac{c_5}{\alpha_X} h^k \quad (8.8.15)$$

can be obtained. On the other hand, $\|\delta\hat{\vartheta}_{gih}\|_X$ satisfies the inequality

$$\|\delta\hat{\vartheta}_{gih}\|_X \leq \|\delta\hat{\vartheta}_{gih}\|_{W^{1,2q_R}(D; \mathbb{R})} \leq c_3 h^k \|\hat{\vartheta}_{gi}\|_{W^{k+1,2q_R}(D; \mathbb{R})} \quad (8.8.16)$$

from Eq. (8.8.11) in which $j = 1$. In Eq. (8.8.16), $\|\hat{\vartheta}_{gi}\|_{W^{k+1,2q_R}(D; \mathbb{R})}$ is bounded.

This is because if Hypothesis 8.8.1 (1) is used in the proof of Theorem 8.6.5, $\hat{\vartheta}_{gi} \in W^{k+1,\infty}(D; \mathbb{R})$ can be obtained. Hence, if Eqs. (8.8.15) and (8.8.16) are substituted into Eq. (8.8.13), the result for the lemma can be obtained. □

Lemma 8.8.4 (Error Estimation of λ_h) *When Hypothesis 8.8.1 is satisfied, there exists a positive constant c_7 which is not dependent on h , and the estimate*

$$\|\delta\lambda_h\|_{\mathbb{R}^{|I_A|}} \leq c_7 h^k$$

holds with respect to λ_h of Eq. (8.8.6). \square

Proof With respect to λ_h of Eq. (8.8.6), the equation

$$\begin{aligned} \delta\lambda_h &= A_h^{-1}(-\delta A_h \lambda + \delta b_h) \\ &= A_h^{-1} \left\{ - \left((\delta g_{ih}, \vartheta_{gj}) \right)_{(i,j) \in I_A^2} + \left((g_i, \delta \vartheta_{gjh}) \right)_{(i,j) \in I_A^2} \right. \\ &\quad \left. + \left((\delta g_{ih}, \vartheta_{g0}) \right)_{i \in I_A} + \left((g_i, \delta \vartheta_{g0h}) \right)_{i \in I_A} \right\} \end{aligned} \quad (8.8.17)$$

is established. If in Eq. (8.8.17), Hypothesis 8.8.1 (4) is used, the estimate

$$\begin{aligned} \|\delta\lambda_h\|_{\mathbb{R}^{|I_A|}} &\leq c_4 \left(1 + |I_A| \max_{i \in I_A} |\lambda_i| \right) \max_{(i,j) \in I_A \times (I_A \cup \{0\})} (|\langle \delta g_{ih}, \vartheta_{gj} \rangle| + |\langle g_i, \delta \vartheta_{gjh} \rangle|) \\ &\quad (8.8.18) \end{aligned}$$

holds. Since $|I_A|$ is bounded, with respect to $|\langle \delta g_{ih}, \vartheta_{gj} \rangle|$ of Eq. (8.8.18), the bound

$$|\langle \delta g_{ih}, \vartheta_{gj} \rangle| \leq c_5 h^k \|\vartheta_{gj}\|_X \quad (8.8.19)$$

is obtained from Lemma 8.8.2. Moreover, with respect to $|\langle g_i, \delta \vartheta_{gjh} \rangle|$,

$$|\langle g_i, \delta \vartheta_{gjh} \rangle| \leq c_6 h^k \|g_i\|_X \quad (8.8.20)$$

can be obtained from Lemma 8.8.3. In Eq. (8.8.20), $\|g_i\|_X$ is bounded. This is because if Hypothesis 8.8.1 (1) is used in the proof of Theorem 8.5.2, $g_i \in W^{k,q_R}(D; \mathbb{R})$ can be obtained. Hence, if Eqs. (8.8.18) and (8.8.19) are substituted into Eq. (8.8.17), the result of the lemma can be obtained. \square

The following results can be obtained based on these lemmas.

Theorem 8.8.5 (Error Estimation of ϑ_g) *When Hypothesis 8.8.1 is satisfied, there exists a positive constant c not dependent on h , and*

$$\|\delta \vartheta_{gh}\|_X \leq c h^k$$

is satisfied with respect to $\delta \vartheta_{gh}$ of Eq. (8.8.7). \square

Proof The following is established based on Eq. (8.8.7):

$$\delta\vartheta_{gh} = \delta\vartheta_{g0h} + \sum_{i \in I_A} (\delta\lambda_{ih}\vartheta_{gi} + \lambda_i \delta\vartheta_{gih}). \quad (8.8.21)$$

From Eq. (8.8.21),

$$\begin{aligned} \|\delta\vartheta_{gh}\|_X &\leq \left(1 + |I_A| \max_{i \in I_A} |\lambda_i|\right) \max_{i \in I_A \cup \{0\}} \|\delta\vartheta_{gih}\|_X \\ &\quad + \|\delta\lambda_h\|_{\mathbb{R}^{|I_A|}} |I_A| \max_{i \in I_A} \|\vartheta_{gi}\|_X \end{aligned} \quad (8.8.22)$$

can be obtained. If Eq. (8.8.22) is substituted with the results of Lemmas 8.8.3 and 8.8.4, the result of the theorem can be obtained. \square

From Theorem 8.8.5, the following can be said with respect to error estimation of finite element solutions with respect to the topology optimization problem of θ -type.

Remark 8.8.6 (Error Estimation of Finite Element Solution ϑ_{gh}) When the numerical solutions of the three boundary value problems (the state determination problem, the adjoint problems and the H^1 gradient method) are obtained by the finite element method with $k = 1$ order basis functions, from Theorem 8.8.5, the error $\|\delta\vartheta_{gh}\|_X$ of search vector ϑ_{gh} reduces to the first order of h with respect to a sequence $\{\mathcal{T}_h\}_{h \rightarrow 0}$ of finite element divisions. \square

8.9 Topology Optimization Problem of Linear Elastic Body

Let us change the state determination problem of θ -type topology optimization problem to a linear elastic problem. Here, a mean compliance minimization problem of a linear elastic body is defined, and let us look at the θ -derivative and second-order θ -derivative. If θ -derivatives and second-order θ -derivatives of the cost functions can be obtained, such a problem can be solved in a similar way to the Poisson problem.

Let D , Γ_D and Γ_N be the domain, Dirichlet boundary and Neumann boundary of a linear elastic problem, similar to the θ -type Poisson problem (Problem 8.2.3). For X and \mathcal{D} , Eqs. (8.1.3) and (8.1.4) respectively will be used.

8.9.1 State Determination Problem

Let us define a linear elastic problem as a state determination problem. Let the linear space U and admissible set \mathcal{S} of state variables be

$$U = \left\{ \mathbf{u} \in H^1(D; \mathbb{R}^d) \mid \mathbf{u} = \mathbf{0}_{\mathbb{R}^d} \text{ on } \Gamma_D \right\}, \quad (8.9.1)$$

$$\mathcal{S} = U \cap W^{1,2q_R}(D; \mathbb{R}^d). \quad (8.9.2)$$

Moreover, Hypothesis 8.2.1 is changed in the following way.

Hypothesis 8.9.1 (Regularity of Known Functions) With respect to $q_R > d$, assume

$$\begin{aligned} \mathbf{b} &\in C^1(X; L^{2q_R}(D; \mathbb{R}^d)), \quad \mathbf{p}_N \in L^{2q_R}(\Gamma_N; \mathbb{R}^d), \\ \mathbf{u}_D &\in H^2(D; \mathbb{R}^d), \quad \mathbf{C} \in L^\infty(D; \mathbb{R}^{d \times d \times d \times d}). \end{aligned} \quad \square$$

On top of this, a problem such as the following is defined with respect to a θ -type linear elastic body such as the one in Fig. 8.12.

Problem 8.9.2 (θ -Type Linear Elastic Problem) Let us suppose that Hypotheses 8.9.1 and 8.2.2 hold. Moreover, let $\alpha > 1$ be a constant and $\phi(\theta)$ is given by Eq. (8.1.1) or Eq. (8.1.2) with respect to $\theta \in \mathcal{D}$. In this case, obtain $\mathbf{u} : D \rightarrow \mathbb{R}^d$ satisfying

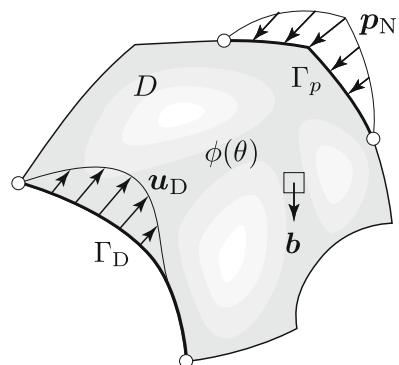
$$-\nabla^\top(\phi^\alpha(\theta) \mathbf{S}(\mathbf{u})) = \mathbf{b}^\top(\theta) \quad \text{in } D, \quad (8.9.3)$$

$$\phi^\alpha(\theta) \mathbf{S}(\mathbf{u}) \mathbf{v} = \mathbf{p}_N \quad \text{on } \Gamma_N, \quad (8.9.4)$$

$$\mathbf{u} = \mathbf{u}_D \quad \text{on } \Gamma_D. \quad (8.9.5)$$

\square

Fig. 8.12 θ -type linear elastic body



For later use, define the Lagrange function with respect to Problem 8.9.2 as

$$\begin{aligned}\mathcal{L}_S(\theta, \mathbf{u}, \mathbf{v}) = & \int_D (-\phi^\alpha(\theta) \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}) + \mathbf{b}(\theta) \cdot \mathbf{v}) \, dx + \int_{\Gamma_N} \mathbf{p}_N \cdot \mathbf{v} \, d\gamma \\ & + \int_{\Gamma_D} \{(\mathbf{u} - \mathbf{u}_D) \cdot (\phi^\alpha(\theta) \mathbf{S}(\mathbf{v}) \mathbf{v}) + \mathbf{v} \cdot (\phi^\alpha(\theta) \mathbf{S}(\mathbf{u}) \mathbf{v})\} \, d\gamma,\end{aligned}$$

where \mathbf{u} is not necessarily the solution of Problem 8.9.2. $\mathbf{v} \in U$ is a Lagrange multiplier. If \mathbf{u} is the solution of Problem 8.9.2,

$$\mathcal{L}_S(\theta, \mathbf{u}, \mathbf{v}) = 0$$

holds with respect to an arbitrary $\mathbf{v} \in U$.

8.9.2 Mean Compliance Minimization Problem

Let us define a topology optimization problem of θ -type with respect to a linear elastic problem. Define the cost function as follows. With respect to the solution \mathbf{u} of Problem 8.9.2,

$$f_0(\theta, \mathbf{u}) = \int_D \mathbf{b}(\theta) \cdot \mathbf{u} \, dx + \int_{\Gamma_N} \mathbf{p}_N \cdot \mathbf{u} \, d\gamma - \int_{\Gamma_D} \mathbf{u}_D \cdot (\phi^\alpha(\theta) \mathbf{S}(\mathbf{u}) \mathbf{v}) \, d\gamma \quad (8.9.6)$$

is referred to as the mean compliance. Moreover, the functional

$$f_1(\theta) = \int_D \phi(\theta) \, dx - c_1 \quad (8.9.7)$$

is referred to as the constraint function with respect to the domain measure of the linear elastic body. Here, c_1 is taken to be a positive constant, such that $f_1(\theta) \leq 0$ holds with respect to some $\theta \in \mathcal{D}$. The reason that the functional f_0 in Eq. (8.9.6) is called mean compliance is as follows. The first and second terms on the right-hand side of Eq. (8.9.6) are the work conducted by the volume force \mathbf{b} and traction \mathbf{p}_N , respectively. Since \mathbf{b} and \mathbf{p}_N are fixed, work conducted here being small means that \mathbf{u} is small. It can be referred to as external work if there are only these two terms. However, the third term on the right-hand side of Eq. (8.9.6) is the negative value of the work done by \mathbf{u}_D . It is because the larger work done by \mathbf{u}_D means that the resistance force with respect to deformation is stronger. Based on these, f_0 will be called mean compliance in the sense that it is the mean value for ease of deformation (compliance).

Using these definitions, the mean compliance minimization problem is defined as follows.

Problem 8.9.3 (Mean Compliance Minimization Problem) Let \mathcal{D} and \mathcal{S} be Eqs. (8.1.4) and (8.9.2), respectively. Let f_0 and f_1 be Eqs. (8.9.6) and (8.9.7), respectively. In this case, obtain θ which satisfies

$$\min_{(\theta, \mathbf{u} - \mathbf{u}_D) \in \mathcal{D} \times \mathcal{S}} \{ f_0(\theta, \mathbf{u}) \mid f_1(\theta) \leq 0, \text{ Problem 8.9.2} \}. \quad \square$$

8.9.3 θ -Derivatives of Cost Functions

Let us obtain the θ -derivative of $f_0(\theta, \mathbf{u})$ using the adjoint variable method. Let the Lagrange function of f_0 be

$$\begin{aligned} \mathcal{L}_0(\theta, \mathbf{u}, \mathbf{v}_0) &= f_0(\theta, \mathbf{u}) + \mathcal{L}_S(\theta, \mathbf{u}, \mathbf{v}_0) \\ &= \int_D \{ -\phi^\alpha(\theta) \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0) + \mathbf{b}(\theta) \cdot (\mathbf{u} + \mathbf{v}_0) \} \, dx \\ &\quad + \int_{\Gamma_N} \mathbf{p}_N \cdot (\mathbf{u} + \mathbf{v}_0) \, d\gamma \\ &\quad + \int_{\Gamma_D} \{ (\mathbf{u} - \mathbf{u}_D) \cdot (\phi^\alpha(\theta) \mathbf{S}(\mathbf{v}_0) \mathbf{v}) \\ &\quad + (\mathbf{v}_0 - \mathbf{u}_D) \cdot (\phi^\alpha(\theta) \mathbf{S}(\mathbf{u}) \mathbf{v}) \} \, d\gamma. \end{aligned} \quad (8.9.8)$$

The Fréchet derivative of \mathcal{L}_0 with respect to arbitrary variation $(\vartheta, \hat{\mathbf{u}}, \hat{\mathbf{v}}_0) \in X \times U \times U$ of $(\theta, \mathbf{u}, \mathbf{v}_0)$ can be written as

$$\begin{aligned} \mathcal{L}'_0(\theta, \mathbf{u}, \mathbf{v}_0) [\vartheta, \hat{\mathbf{u}}, \hat{\mathbf{v}}_0] &= \mathcal{L}_{0\theta}(\theta, \mathbf{u}, \mathbf{v}_0) [\vartheta] + \mathcal{L}_{0\mathbf{u}}(\theta, \mathbf{u}, \mathbf{v}_0) [\hat{\mathbf{u}}] + \mathcal{L}_{0\mathbf{v}_0}(\theta, \mathbf{u}, \mathbf{v}_0) [\hat{\mathbf{v}}_0]. \end{aligned} \quad (8.9.9)$$

Each term is considered below.

The third term on the right-hand side of Eq. (8.9.9) becomes

$$\mathcal{L}_{0\mathbf{v}_0}(\theta, \mathbf{u}, \mathbf{v}_0) [\hat{\mathbf{v}}_0] = \mathcal{L}_{S\mathbf{v}_0}(\theta, \mathbf{u}, \mathbf{v}_0) [\hat{\mathbf{v}}_0] = \mathcal{L}_S(\theta, \mathbf{u}, \hat{\mathbf{v}}_0). \quad (8.9.10)$$

Equation (8.9.10) is a Lagrange function of the state determination problem (Problem 8.9.2). Hence, if \mathbf{u} is the weak solution of the state determination problem, the third term of the right-hand side of Eq. (8.9.9) is zero.

Moreover, the second term on the right-hand side of Eq. (8.9.9) becomes

$$\begin{aligned}
 \mathcal{L}_{0u}(\theta, \mathbf{u}, \mathbf{v}_0)[\hat{\mathbf{u}}] &= \int_D (-\phi^\alpha(\theta) \mathbf{S}(\hat{\mathbf{u}}) \cdot \mathbf{E}(\mathbf{v}_0) + \mathbf{b}(\theta) \cdot \hat{\mathbf{u}}) \, dx + \int_{\Gamma_N} \mathbf{p}_N \cdot \hat{\mathbf{u}} \, d\gamma \\
 &\quad + \int_{\Gamma_D} \{\hat{\mathbf{u}} \cdot (\phi^\alpha(\theta) \mathbf{S}(\mathbf{v}_0) \mathbf{v}) + (\mathbf{v}_0 - \mathbf{u}_D) \cdot (\phi^\alpha(\theta) \mathbf{S}(\hat{\mathbf{u}}) \mathbf{v})\} \, d\gamma \\
 &= \mathcal{L}_S(\theta, \mathbf{v}_0, \hat{\mathbf{u}}). \tag{8.9.11}
 \end{aligned}$$

Here, if \mathbf{v}_0 is chosen so that Eq. (8.9.11) becomes zero, the second term on the right-hand side of Eq. (8.9.9) vanishes. This relationship shows that the self-adjoint relationship

$$\mathbf{u} = \mathbf{v}_0 \tag{8.9.12}$$

holds.

Furthermore, the first-term on the right-hand side of Eq. (8.9.9) becomes

$$\mathcal{L}_{0\theta}(\theta, \mathbf{u}, \mathbf{v}_0)[\vartheta] = \int_D \left\{ \mathbf{b}' \cdot (\mathbf{u} + \mathbf{v}_0) - \alpha \phi^{\alpha-1} \phi' \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0) \right\} \vartheta \, dx. \tag{8.9.13}$$

Hence, \mathbf{u} is taken to be a weak solution of Problem 8.9.2 and the self-adjoint relationship (Eq. (8.9.12)) is assumed to hold. If $f_0(\theta, \mathbf{u})$ in this case is denoted as $\tilde{f}_0(\theta)$, we can write

$$\tilde{f}'_0(\theta)[\vartheta] = \mathcal{L}_{0\theta}(\theta, \mathbf{u}, \mathbf{v}_0)[\vartheta] = \langle g_0, \vartheta \rangle, \tag{8.9.14}$$

where

$$g_0 = 2\mathbf{b}' \cdot \mathbf{u} - \alpha \phi^{\alpha-1} \phi' \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{u}). \tag{8.9.15}$$

On the other hand, with respect to $f_1(\theta)$,

$$f'_1(\theta)[\vartheta] = \int_D \phi' \vartheta \, dx = \langle g_1, \vartheta \rangle \tag{8.9.16}$$

is established with respect to an arbitrary $\vartheta \in X$.

Based on the above results, the function space which contains g_0 of Eq. (8.9.15) becomes the same result as Theorem 8.5.2. Hence, by applying H^1 gradient method, the fact that the search vector ϑ_g is in class $C^{0,1}$ is guaranteed.

8.9.4 Second-Order θ -Derivatives of Cost Functions

Furthermore, the second-order θ -derivatives of mean compliance f_0 and the constraint cost function f_1 with respect to the domain measure of linear elastic body can also be obtained. Here, we will follow the procedure shown in Sect. 8.5.2.

Firstly, let us think about the second-order θ -derivative of f_0 . To correspond to Hypothesis 8.5.4 (1), here \mathbf{b} is assumed not to be a function of θ . The relationship corresponding to Hypothesis 8.5.4 (2) is satisfied here.

The Lagrange function \mathcal{L}_0 of f_0 is defined by Eq. (8.9.8). Viewing (θ, \mathbf{u}) as a design variable, its admissible set and admissible set of directions are set as

$$S = \{(\theta, \mathbf{u}) \in \mathcal{D} \times \mathcal{S} \mid \mathcal{L}_S(\theta, \mathbf{u}, \mathbf{v}) = 0 \text{ for all } \mathbf{v} \in U\},$$

$$T_S(\theta, \mathbf{u}) = \{(\vartheta, \hat{\mathbf{v}}) \in X \times U \mid \mathcal{L}_{S\theta\mathbf{u}}(\theta, \mathbf{u}, \mathbf{v})[\vartheta, \hat{\mathbf{v}}] = 0 \text{ for all } \mathbf{v} \in U\}.$$

The second-order Fréchet partial derivative of \mathcal{L}_0 of Eq. (8.9.8) with respect to arbitrary variations $(\vartheta_1, \hat{\mathbf{v}}_1), (\vartheta_2, \hat{\mathbf{v}}_2) \in T_S(\theta, \mathbf{u})$ of design variable $(\theta, \mathbf{u}) \in S$ becomes

$$\begin{aligned} & \mathcal{L}_{0(\theta, \mathbf{u})(\theta, \mathbf{u})}(\theta, \mathbf{u}, \mathbf{v}_0)[(\vartheta_1, \hat{\mathbf{v}}_1), (\vartheta_2, \hat{\mathbf{v}}_2)] \\ &= \mathcal{L}_{0\theta\theta}(\theta, \mathbf{u}, \mathbf{v}_0)[\vartheta_1, \vartheta_2] + \mathcal{L}_{0\theta\mathbf{u}}(\theta, \mathbf{u}, \mathbf{v}_0)[\vartheta_1, \hat{\mathbf{v}}_2] \\ & \quad + \mathcal{L}_{0\theta\mathbf{u}}(\theta, \mathbf{u}, \mathbf{v}_0)[\vartheta_2, \hat{\mathbf{v}}_1] + \mathcal{L}_{0\mathbf{u}\mathbf{u}}(\theta, \mathbf{u}, \mathbf{v}_0)[\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2]. \end{aligned} \quad (8.9.17)$$

Each term on the right-hand side of Eq. (8.9.17) becomes

$$\mathcal{L}_{0\theta\theta}(\theta, \mathbf{u}, \mathbf{v}_0)[\vartheta_1, \vartheta_2] = \int_D -(\phi^\alpha(\theta))'' \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0) \vartheta_1 \vartheta_2 \, dx, \quad (8.9.18)$$

$$\mathcal{L}_{0\theta\mathbf{u}}(\theta, \mathbf{u}, \mathbf{v}_0)[\vartheta_1, \hat{\mathbf{v}}_2] = \int_D -(\phi^\alpha(\theta))' \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0) \vartheta_1 \, dx, \quad (8.9.19)$$

$$\mathcal{L}_{0\theta\mathbf{u}}(\theta, \mathbf{u}, \mathbf{v}_0)[\vartheta_2, \hat{\mathbf{v}}_1] = \int_D -(\phi^\alpha(\theta))' \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0) \vartheta_2 \, dx, \quad (8.9.20)$$

$$\mathcal{L}_{0\mathbf{u}\mathbf{u}}(\theta, \mathbf{u}, \mathbf{v}_0)[\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2] = 0. \quad (8.9.21)$$

Here, $\mathbf{u} - \mathbf{u}_D, \mathbf{v}_0 - \mathbf{u}_D, \hat{\mathbf{v}}_1$ and $\hat{\mathbf{v}}_2$ use the fact that $\mathbf{0}_{\mathbb{R}^d}$ on Γ_D . Moreover, $(\phi^\alpha(\theta))'$ and $(\phi^\alpha(\theta))''$ are Eqs. (8.5.14) and (8.5.15), respectively. Here, with respect to an arbitrary variation $(\vartheta_j, \hat{\mathbf{v}}_j) \in T_S(\theta, \mathbf{u})$ for $j \in \{1, 2\}$, the Fréchet partial derivative of Lagrange function \mathcal{L}_S of the state determination problem becomes

$$\begin{aligned} & \mathcal{L}_{S\theta\mathbf{u}}(\theta, \mathbf{u}, \mathbf{v})[\vartheta_j, \hat{\mathbf{v}}_j] \\ &= \int_D \left\{ -(\phi^\alpha(\theta))' \vartheta_j \mathbf{S}(\mathbf{v}) - \phi^\alpha(\theta) \mathbf{S}(\hat{\mathbf{v}}_j) \right\} \cdot \mathbf{E}(\mathbf{v}) \, dx \\ &= 0 \end{aligned} \quad (8.9.22)$$

with respect to an arbitrary $\mathbf{v} \in U$. Here, the fact that \mathbf{v} and $\hat{\mathbf{v}}_j$ are $\mathbf{0}_{\mathbb{R}^d}$ on Γ_D is used. From Eq. (8.9.22), the equation

$$\mathbf{S}(\hat{\mathbf{v}}_j) = -\frac{(\phi^\alpha(\theta))'}{\phi^\alpha(\theta)} \vartheta_j \mathbf{S}(\mathbf{u}) \quad \text{in } D \quad (8.9.23)$$

can be obtained. Hence, substituting $\hat{\mathbf{v}}_j$ of Eq. (8.9.23) into $\hat{\mathbf{v}}_1$ in Eq. (8.9.20) and $\hat{\mathbf{u}}_2$ in Eq. (8.9.19), and considering Dirichlet boundary conditions in the state determination problem and the adjoint problems with respect to f_1, \dots, f_m as well as $\hat{\mathbf{v}}_j = \mathbf{0}_{\mathbb{R}^d}$ on Γ_D , the equations

$$\begin{aligned} \mathcal{L}_{0\theta\mathbf{u}}(\theta, \mathbf{u}, \mathbf{v}_0)[\vartheta_1, \hat{\mathbf{v}}_2] &= \mathcal{L}_{0\theta\mathbf{u}}(\theta, \mathbf{u}, \mathbf{v}_0)[\vartheta_2, \hat{\mathbf{v}}_1] \\ &= \int_D \frac{(\phi^\alpha(\theta))^2}{\phi^\alpha(\theta)} \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0) \vartheta_1 \vartheta_2 \, dx \end{aligned} \quad (8.9.24)$$

is obtained.

Summarizing the results above, by substituting Eqs. (8.9.24) and (8.9.18) into Eq. (8.9.17), the second-order θ -derivative of mean compliance f_0 becomes

$$\begin{aligned} h_0(\theta, \mathbf{u}, \mathbf{v}_0)[\vartheta_1, \vartheta_2] &= \int_D \left\{ 2 \frac{(\phi^\alpha(\theta))^2}{\phi^\alpha(\theta)} - (\phi^\alpha(\theta))'' \right\} \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0) \vartheta_1 \vartheta_2 \, dx \\ &= \int_D \beta(\alpha, \theta) \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0) \vartheta_1 \vartheta_2 \, dx, \end{aligned} \quad (8.9.25)$$

where $\beta(\alpha, \theta)$ is given by Eq. (8.5.20). Furthermore, if a self-adjoint relationship is used, $\mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0) > 0$ and $h_0(\theta, \mathbf{u}, \mathbf{v}_0)[\cdot, \cdot]$ becomes a coercive and bounded bilinear form on X .

On the other hand, the second-order θ -derivative of $f_1(\theta)$ becomes

$$h_1(\theta)[\vartheta_1, \vartheta_2] = f_1''(\theta)[\vartheta_1, \vartheta_2] = \int_D \phi''(\theta) \vartheta_1 \vartheta_2 \, dx \quad (8.9.26)$$

with respect to an arbitrary $\vartheta_1, \vartheta_2 \in X$. Here, when $\phi(\theta)$ of Eq. (8.1.1) is used, we get

$$\phi''(\theta) = -\frac{1}{\pi} \frac{2\theta}{(1+\theta^2)^2}. \quad (8.9.27)$$

Moreover, when $\phi(\theta)$ is given by Eq. (8.1.2), the equation

$$\phi''(\theta) = -\operatorname{sech}^2 \theta \tanh \theta \quad (8.9.28)$$

holds. Figure 8.13 shows the graphs of $\phi''(\theta)$.

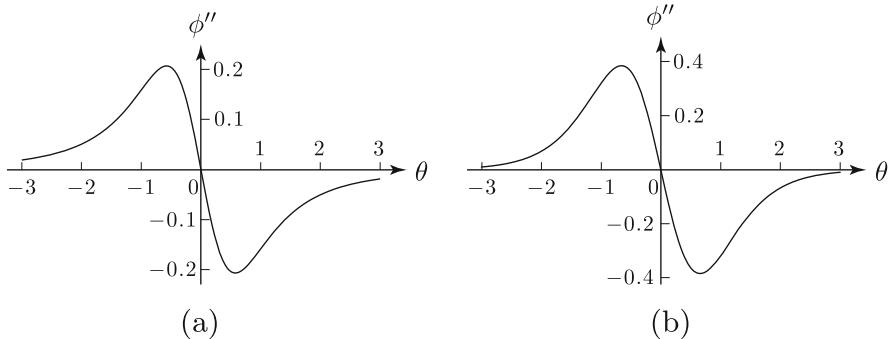


Fig. 8.13 Coefficient functions $\phi''(\theta)$ in h_1 . (a) $\phi(\theta) = \tan^{-1}\theta/\pi + 1/2$. (b) $\phi(\theta) = (\tanh\theta + 1)/2$

In this way, in the mean compliance minimization problem, the second-order θ -derivative of the object cost function f_0 is coercive but the second-order θ -derivative of the constraint function f_1 is not. Hence, if using the Newton method (Problem 8.6.6), an additional term for capturing coerciveness becomes necessary.

8.9.5 Second-Order θ -Derivative of Cost Function Using Lagrange Multiplier Method

When the Lagrange multiplier method is used to obtain the second-order θ -derivative of the mean compliance f_0 , it becomes as follows. Fixing ϑ_1 , we define the Lagrange function for $\tilde{f}'_0(\theta) [\vartheta_1] = \langle g_0, \vartheta_1 \rangle$ in Eq. (8.9.14) by

$$\mathcal{L}_{10}(\theta, \mathbf{u}, \mathbf{w}_0) = \langle g_0, \vartheta_1 \rangle + \mathcal{L}_S(\theta, \mathbf{u}, \mathbf{w}_0), \quad (8.9.29)$$

where \mathcal{L}_S is the Lagrange function of Problem 8.9.2, and $\mathbf{w}_0 \in U$ is the adjoint variable provided for \mathbf{u} in \mathbf{g}_0 .

With respect to arbitrary variations $(\vartheta_2, \hat{\mathbf{u}}, \hat{\mathbf{w}}_0) \in X \times U^2$ of $(\theta, \mathbf{u}, \mathbf{w}_0)$, the Fréchet derivative of \mathcal{L}_{10} is written as

$$\begin{aligned} \mathcal{L}'_{10}(\theta, \mathbf{u}, \mathbf{w}_0) [\vartheta_2, \hat{\mathbf{u}}, \hat{\mathbf{w}}_0] &= \mathcal{L}_{10\theta}(\theta, \mathbf{u}, \mathbf{w}_0) [\vartheta_2] + \mathcal{L}_{10\mathbf{u}}(\theta, \mathbf{u}, \mathbf{w}_0) [\hat{\mathbf{u}}] \\ &\quad + \mathcal{L}_{10\mathbf{w}_0}(\theta, \mathbf{u}, \mathbf{w}_0) [\hat{\mathbf{w}}_0]. \end{aligned} \quad (8.9.30)$$

The third term on the right-hand side of Eq. (8.9.30) vanishes if \mathbf{u} is the solution of the state determination problem.

The second term on the right-hand side of Eq. (8.9.30) is

$$\begin{aligned} & \mathcal{L}_{10u}(\theta, \mathbf{u}, \mathbf{w}_0)[\hat{\mathbf{u}}] \\ &= \int_D \left\{ \left(2\mathbf{b}' \cdot \hat{\mathbf{u}} - 2\alpha\phi^{\alpha-1}\phi' \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\hat{\mathbf{u}}) \right) \vartheta_1 - \phi^\alpha \mathbf{S}(\mathbf{w}_0) \cdot \mathbf{E}(\hat{\mathbf{u}}) \right\} dx. \end{aligned} \quad (8.9.31)$$

Here, the condition that Eq. (8.9.31) is zero for arbitrary $\hat{\mathbf{u}} \in U$ is equivalent to setting \mathbf{w}_0 to be the solution of the following adjoint problem.

Problem 8.9.4 (Adjoint Problem of \mathbf{w}_0 with Respect to (g_0, ϑ_1)) Under the assumption of Problem 8.9.2, let $\vartheta_1 \in X$ be given. Find $\mathbf{w}_0 = \mathbf{w}_0(\vartheta_1) \in U$ satisfying

$$\begin{aligned} -\nabla^\top(\phi^\alpha \mathbf{S}(\mathbf{w}_0)) &= 2 \left(\nabla^\top \left(\alpha\phi^{\alpha-1}\phi' \mathbf{S}(\mathbf{u}) \right) + \mathbf{b}'^\top \right) \vartheta_1 \quad \text{in } D, \\ \phi^\alpha \mathbf{S}(\mathbf{w}_0) \mathbf{v} &= 2\alpha\phi^{\alpha-1}\phi' \mathbf{S}(\mathbf{u}) \mathbf{v} \vartheta_1 \quad \text{on } \Gamma_N, \\ \mathbf{w}_0 &= \mathbf{0}_{\mathbb{R}^d} \quad \text{on } \Gamma_D. \end{aligned}$$

□

Finally, the first term on the right-hand side of Eq. (8.9.30) becomes

$$\begin{aligned} & \mathcal{L}_{10\theta}(\theta, \mathbf{u}, \mathbf{w}_0)[\vartheta_2] \\ &= \int_D \left[\left\{ 2\mathbf{b}'' \cdot \mathbf{u} - \left(\alpha(\alpha-1)\phi^{\alpha-2}\phi'^2 + \alpha\phi^{\alpha-1}\phi'' \right) \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{u}) \right\} \vartheta_1 \right. \\ & \quad \left. - \alpha\phi^{\alpha-1}\phi' \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{w}_0) + \mathbf{b}' \cdot \mathbf{w}_0 \right] \vartheta_2 dx. \end{aligned}$$

Here, \mathbf{u} and $\mathbf{w}_0(\vartheta_1)$ are assumed to be the weak solutions of Problems 8.9.2 and 8.9.4, respectively. If we denote $f_i(\theta, \mathbf{u})$ by $\tilde{f}_i(\theta)$, we have the relation:

$$\begin{aligned} \mathcal{L}_{10\theta}(\theta, \mathbf{u}, \mathbf{v}_0, \mathbf{w}_0(\vartheta_1), \mathbf{z}_0(\vartheta_1))[\vartheta_2] &= \tilde{f}_0''(\theta)[\vartheta_1, \vartheta_2] \\ &= \langle g_{H0}(\theta, \vartheta_1), \vartheta_2 \rangle, \end{aligned} \quad (8.9.32)$$

where the Hesse gradient g_{H0} of the mean compliance is given by

$$\begin{aligned} & g_{H0}(\theta, \vartheta_1) \\ &= \left\{ - \left(\alpha(\alpha-1)\phi^{\alpha-2}\phi'^2 + \alpha\phi^{\alpha-1}\phi'' \right) \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{u}) + 2\mathbf{b}'' \cdot \mathbf{u} \right\} \vartheta_1 \\ & \quad - \alpha\phi^{\alpha-1}\phi' \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{w}_0(\vartheta_1)) + \mathbf{b}' \cdot \mathbf{w}_0(\vartheta_1). \end{aligned} \quad (8.9.33)$$

If \mathbf{b} is not a function of θ , with respect to the solution \mathbf{w}_0 of Problem 8.9.4,

$$\mathbf{E}(\mathbf{w}_0(\vartheta_1)) = -2 \frac{\alpha\phi'}{\phi} \mathbf{E}(\mathbf{v}_0) \vartheta_1 \quad (8.9.34)$$

holds. Substituting Eq. (8.9.34) into Eq. (8.9.33), it can be confirmed that Eq. (8.9.32) accords with Eq. (8.9.25).

8.9.6 Numerical Example

The results of mean compliance minimization for a two-dimensional linear elastic body with a boundary condition referred to as the coat-hanging problem is shown in Figs. 8.14, 8.15, 8.16. The initial density ($\theta = 0$) and the boundary condition for the linear elastic problem are shown in Fig. 8.14a. A domain in which the density is constrained ($\bar{\Omega}_{C0}$ in Eq. (8.1.3)) was not set. The program is written using the programming language FreeFEM (<https://freefem.org/>) [66] using the finite element method. In the finite element analyses of the linear elastic problem and the H^1 gradient method or the H^1 Newton method, the second-order triangular elements were used. In the case using the H^1 Newton method, the routine of the H^1 Newton method was started from $k_N = 10$. The results were changed with c_a in Eq. (8.7.1),

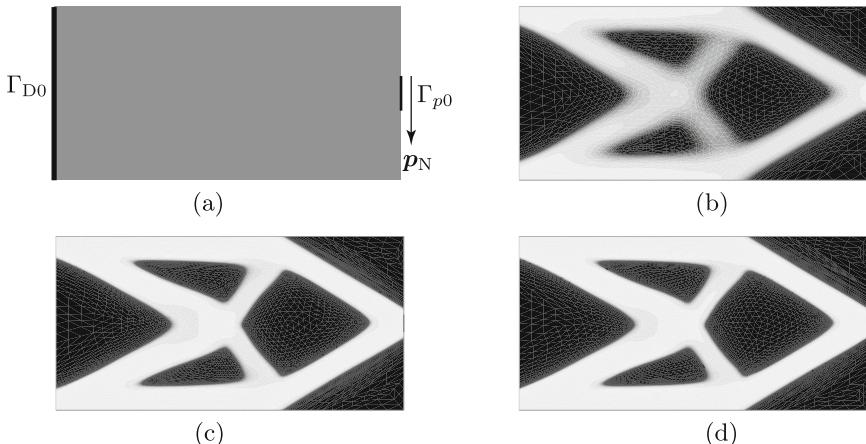


Fig. 8.14 Numerical example of mean compliance minimization: density (a) Initial density and boundary condition. (b) H^1 gradient method ($k = 100$). (c) H^1 Newton method ($k = 100$). (d) H^1 Newton method (Hesse gradient, $k = 100$)

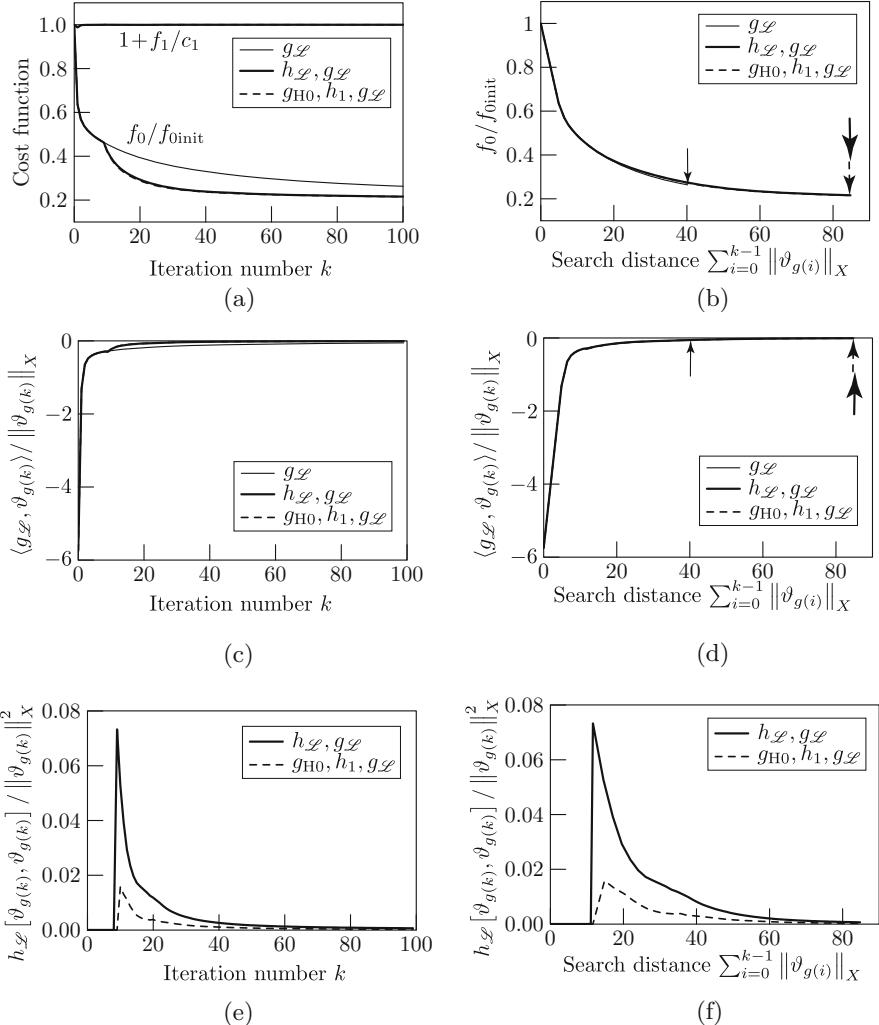


Fig. 8.15 Numerical example of mean compliance minimization: cost functions and gradients and Hessians of f_0 on the search path ($g_{\mathcal{L}}$: H¹ gradient method, $h_{\mathcal{L}}, g_{\mathcal{L}}$: H¹ Newton method, $g_{H0}, h_1, g_{\mathcal{L}}$: H¹ Newton method using the Hesse gradient). (a) Cost functions. (b) Cost functions (search distance). (c) Gradient of f_0 on the search path. (d) Gradient of f_0 on the search path (search distance). (e) Hessian of f_0 on the search path. (f) Hessian of f_0 on the search path (search distance)

c_D in Eq. (8.6.3), c_{D1} and c_{D0} in Eq. (8.6.7), c_h in Eq. (8.7.8) and the parameter (`erre1as`) to control the error level in the adaptive mesh. For details, we refer the readers to the programs themselves.¹

¹See Electronic supplementary material.

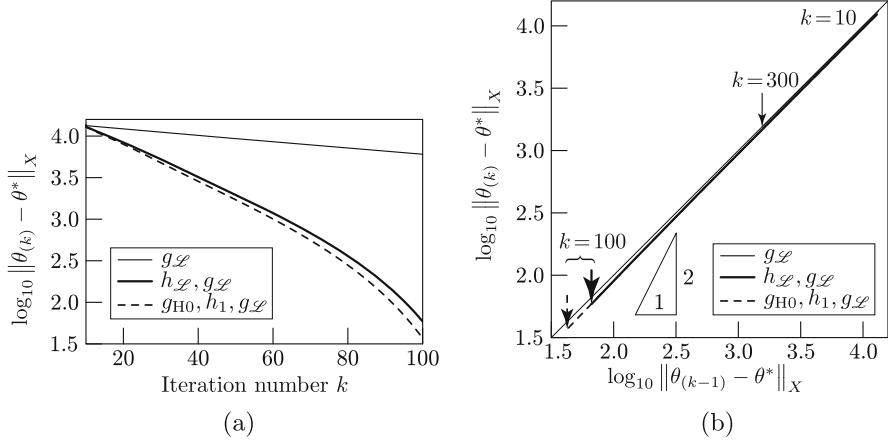


Fig. 8.16 Numerical example of mean compliance minimization problem: distance $\|\theta_k - \theta^*\|_X$ from an approximate minimum point θ^* ($g_{\mathcal{L}}$: the gradient method, $h_{\mathcal{L}}, g_{\mathcal{L}}$: the Newton method, $g_{H0}, h_1, g_{\mathcal{L}}$: the Newton method using the Hesse gradient). **(a)** Iteration history. **(b)** $(k-1)$ -th vs. k -th plot

Figure 8.14b-d show the densities obtained by the three methods (H^1 gradient method using $g_{\mathcal{L}} = g_0 + \lambda_1 g_1$, H^1 Newton method using $h_{\mathcal{L}} = h_0 + \lambda_1 h_1$ and $g_{\mathcal{L}}$, and H^1 Newton method using $g_{H0}, h_1, g_{\mathcal{L}}$). Figure 8.15a shows the cost functions $f_0/f_0\text{init}$ and $1 + f_1/c_1$ normalized with f_0 at the initial density denoted by $f_0\text{init}$ and c_1 set with the domain integral of the initial density (volume), respectively, with respect to the iteration number k . Figure 8.15b shows those values with respect to the distance $\sum_{i=0}^{k-1} \|\vartheta_{g(i)}\|_X$ on the search path in X . The graphs of f_0 's gradient (the gradient of the Lagrange function $\mathcal{L} = \mathcal{L}_0 + \lambda_1 f_1$) calculated as $\langle g_{\mathcal{L}}, \vartheta_{g(k)} \rangle / \|\vartheta_{g(k)}\|_X$ are shown in Fig. 8.15c,d with respect to the iteration number and the search distance, respectively. Moreover, Fig. 8.15e,f show the graphs of f_0 's second-order derivative $h_{\mathcal{L}} [\vartheta_{g(k)}, \vartheta_{g(k)}] / \|\vartheta_{g(k)}\|_X^2$ (in the case of the Newton method using the Hesse gradient, $((g_{H0}, \vartheta_{g(k)}) + \lambda_1 h_1 [\vartheta_{g(k)}, \vartheta_{g(k)}]) / \|\vartheta_{g(k)}\|_X^2$) with respect to the iteration number and the search distance, respectively. In these notations, the norm of the i -th search vector is defined by

$$\|\vartheta_{g(i)}\|_X = \left(\int_D \left(\nabla \vartheta_{g(i)} \cdot \nabla \vartheta_{g(i)} + \vartheta_{g(i)}^2 \right) dx \right)^{1/2}. \quad (8.9.35)$$

The computational times until $k = 100$ by PC were 6.897, 17.073 and 32.394 s by the H^1 gradient method, the H^1 Newton method and the H^1 Newton method using the Hesse gradient, respectively.

We explain the numerical results and give some considerations as follows. The graphs in Fig. 8.15a clearly show that the convergence speed with respect to the iteration number k is faster when using the H^1 Newton method than when applying

the H^1 gradient method. However, when the H^1 Newton method started, c_{D1} and c_{D0} in Eq. (8.6.7) were replaced with smaller values so as to avoid any numerical instability during iterations. As a result, it can be considered that the convergence speed was increased by enlarging the step size. In reality, the following phenomenon was observed. When we set c_h to zero (H^1 gradient method), the computation fails at $k = k_N$. However, when we put a larger value for c_{D0} in Eq. (8.6.7), we have a convergence similar to the H^1 Newton method. Moreover, based on the fact that the three graphs plotted in Fig. 8.15b coincide, we conclude that the search paths due from each methods are actually the same. Based on these observations, we infer that the H^1 Newton method is superior to the first-order method, in the sense that we can take larger values for the step size.

Based from the results shown in Fig. 8.15d,f, we also draw some particular observations around the minimum point as follows. Firstly, since the Hessian of f_0 on the search path has a positive value, we deduce that the point of convergence is a local minimum. Secondly, we noticed that both the first derivative and the Hessian eventually diminish to zero. This key finding is observed because we assumed a sigmoid function for the density of the design valuable θ , and obtained a small variation of the density around the minimum point where the density converges to 0 or 1 that causes a small variation of the cost functions.

In addition, Fig. 8.16a shows the graphs of the distance $\|\theta_{(k)} - \theta^*\|_X$ from an approximate minimum point θ^* obtained by the three methods with respect to the iteration number k . The approximate minimum point θ^* was given by the numerical solution of θ when the iteration time is taken larger than the given value in the H^1 Newton method. From this figure, it can be confirmed that the convergence orders for the results by the H^1 Newton methods are more than the first order. However, Fig. 8.16b, plotting the k -th distance $\|\theta_k - \theta^*\|_X$ with respect to the $(k - 1)$ -th distance (the gradient of the graph shows the order of convergence as explained by using Eq. (3.8.13)), shows that the convergence order of the H^1 Newton method is less than the second order, while more than the first order. This result is possibly due to the fact that the bilinear form a_X in X was added to the original Hessian in order to ensure coerciveness and boundedness of the left-hand side of Eq. (8.6.6). By this addition, the H^1 Newton method has a different structure from the original Newton method. It is still unclear, however, whether a solution with second-order convergence exists for the problem analyzed in this section.

8.10 Topology Optimization Problem of Stokes Flow Field

The fact that the topology optimization problem can be constructed even with respect to a flow field is shown in the literature [1, 28, 49, 53, 57, 58, 131]. Here, the mean flow resistance minimization problem of the one-dimensional branched Stokes flow field mentioned in Sect. 1.3 is extended to a $d \in \{2, 3\}$ -dimensional topology optimization problem of θ -type. Here, θ -derivatives of cost functions and

second-order θ -derivatives are shown up as far as can be sought. In this section, D is assumed to be a Stokes flow field and X and \mathcal{D} are taken to be defined in Eqs. (8.1.3) and (8.1.4), respectively.

8.10.1 State Determination Problem

Let us define a Stokes problem as a state determination problem. A Stokes problem (Problem 5.5.1) was defined in Sect. 5.5 but here a Stokes flow field of θ -type such as the one in Fig. 8.17 is considered. In this regard, some definitions will be added. With respect to U and \mathcal{S} , Eqs. (8.9.1) and (8.9.2) will be used respectively, but $\Gamma_D = \partial D$. Furthermore, with respect to $q_R > d$, put

$$P = \left\{ q \in L^2(D; \mathbb{R}) \mid \int_D q \, dx = 0 \right\}, \quad (8.10.1)$$

$$\mathcal{Q} = P \cap L^{2q_R}(D; \mathbb{R}). \quad (8.10.2)$$

In a topology optimization problem of a flow field, a flow field passing through porous media (penetration flow) is used. In a penetration flow, between the flow speed \mathbf{u} and pressure p , the Darcy law given by

$$\mathbf{u} = -\frac{k}{\mu} \nabla p$$

is assumed to hold. Here, k and μ are positive constants known as penetration and viscosity coefficients. In a topology optimization problem of a flow field, replace the constant μ/k representing the difficulty of penetration with

$$\psi(\phi) = \psi_1 \left\{ 1 - \frac{\phi(1+\alpha)}{\phi+\alpha} \right\} = \psi_1 \frac{\alpha(1-\phi)}{\alpha+\phi} \quad (8.10.3)$$

and ∇p in the Stokes equation with $\psi(\phi(\theta)) \mathbf{u} + \nabla p$. Here, ϕ represents the fluid content equivalent to the density of fluid and its range is assumed to be limited to $[0, 1]$. Hence, similarly to the density in the previous sections, ϕ is assumed to be

Fig. 8.17 Stokes flow field of θ -type

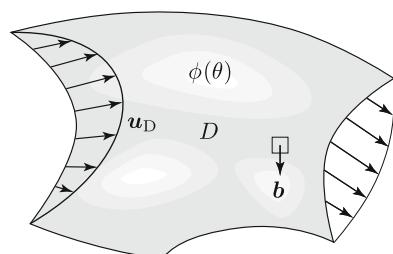
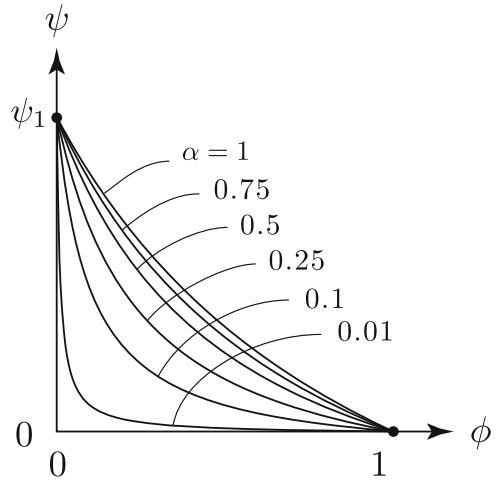


Fig. 8.18 Coefficient ψ expressing the flow resistance with respect to fluid content ϕ



given by the sigmoid function with respect to the design variable $\theta \in X$. Moreover, ψ_1 is a positive constant which gives the maximum value of the resistance to flow, α is a constant controlling the non-linearity and is chosen from $(0, 1]$. In the paper [1], it is changed from 0.01 to 1 along with calculation progression. Figure 8.18 shows the function $\psi(\phi)$.

Here, the following assumption is set.

Hypothesis 8.10.1 (Regularities of Known Functions) With respect to $q_R > d$,

$$\mathbf{b} \in L^{2q_R}(D; \mathbb{R}^d), \quad \mathbf{u}_D \in \left\{ \mathbf{u} \in W^{1,2q_R}(D; \mathbb{R}^d) \mid \nabla \cdot \mathbf{u} = 0 \text{ in } D \right\}$$

is assumed. \square

Using these assumptions, a Stokes problem of θ -type is defined in the following way.

Problem 8.10.2 (θ -Type Stokes Problem) Let \mathbf{b} and \mathbf{u}_D satisfy Hypothesis 8.10.1, and Hypothesis 8.2.2 holds with respect to the opening angles of boundary corner points. Moreover, $\psi(\phi)$ is taken to be Eq. (8.10.3). Furthermore, $\phi(\theta)$ is assumed to be given by Eq. (8.1.1) or Eq. (8.1.2). Here, obtain $(\mathbf{u}, p) : D \rightarrow \mathbb{R}^{d+1}$ which satisfies

$$-\nabla^\top (\mu \nabla \mathbf{u}^\top) + \psi(\phi(\theta)) \mathbf{u}^\top + \nabla^\top p = \mathbf{b}^\top \quad \text{in } D, \quad (8.10.4)$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } D, \quad (8.10.5)$$

$$\mathbf{u} = \mathbf{u}_D \quad \text{on } \partial D, \quad (8.10.6)$$

$$\int_D p \, dx = 0. \quad (8.10.7)$$

\square

For later use, the Lagrange function with respect to Problem 8.10.2 is defined as

$$\begin{aligned} \mathcal{L}_S(\theta, \mathbf{u}, p, \mathbf{v}, q) &= \int_D \left\{ -\mu \left(\nabla \mathbf{u}^\top \right) \cdot \left(\nabla \mathbf{v}^\top \right) - \psi(\phi(\theta)) \mathbf{u} \cdot \mathbf{v} \right. \\ &\quad \left. + p \nabla \cdot \mathbf{v} + \mathbf{b} \cdot \mathbf{v} + q \nabla \cdot \mathbf{u} \right\} dx \\ &\quad + \int_{\partial D} \{ (\mathbf{u} - \mathbf{u}_D) \cdot (\mu \partial_\nu \mathbf{v} - q \mathbf{v}) + \mathbf{v} \cdot (\mu \partial_\nu \mathbf{u} - p \mathbf{v}) \} d\gamma, \end{aligned} \quad (8.10.8)$$

where (\mathbf{u}, p) is not necessarily the solution of Problem 8.10.2, and $(\mathbf{v}, q) \in U \times P$ is a Lagrange multiplier.

When (\mathbf{u}, p) is a solution to Problem 8.10.2, the equation

$$\mathcal{L}_S(\theta, \mathbf{u}, p, \mathbf{v}, q) = 0$$

is established with respect to an arbitrary $(\mathbf{v}, q) \in U \times P$.

8.10.2 Mean Flow Resistant Minimization Problem

Let us define a topology optimization problem of θ -type with respect to a Stokes flow field. Define the cost functions as follows. Firstly, let us define a cost function representing the resistance to flow as

$$f_0(\theta, \mathbf{u}, p) = - \int_D \mathbf{b} \cdot \mathbf{u} dx + \int_{\partial D} \mathbf{u}_D \cdot (\mu \partial_\nu \mathbf{u} - p \mathbf{v}) d\gamma. \quad (8.10.9)$$

The first term on the right-hand side of Eq. (8.10.9) represents the negative value of the rate of work due to the volume force. This value was given a negative sign because the greater it is, the greater the flow speed is. On the other hand, the second term on the right-hand side of Eq. (8.10.9) is equivalent to the energy per unit time lost through the viscosity inside Stokes flow field represented by the boundary integral. From the fact that these express the property of flow resistance, f_0 will be referred to as the mean flow resistance. With respect to this,

$$f_1(\theta) = \int_D \phi(\theta) dx - c_1 \quad (8.10.10)$$

is the cost function for constraint with respect to the domain measure of the flow field. Here, c_1 is a positive constant such that $f_1(\theta) \leq 0$ holds with respect to some $\theta \in \mathcal{D}$. Using these, the minimization problem of mean flow resistance is defined as follows.

Problem 8.10.3 (Mean Flow Resistance Minimization Problem) Let \mathcal{D} , \mathcal{S} and \mathcal{Q} be Eqs. (8.1.4), (8.9.2) and (8.10.2), respectively. Let (\mathbf{u}, p) be the solution of Problem 8.10.2 with respect to $\theta \in \mathcal{D}$ and f_0 and f_1 are given by Eqs. (8.10.9) and (8.10.10). In this case, obtain θ which satisfies

$$\min_{(\theta, \mathbf{u} - \mathbf{u}_D, p) \in \mathcal{D} \times \mathcal{S} \times \mathcal{Q}} \{ f_0(\theta, \mathbf{u}, p) \mid f_1(\theta) \leq 0, \text{ Problem 8.10.2} \}. \quad \square$$

8.10.3 θ -Derivatives of Cost Functions

Let us obtain the θ -derivative of $f_0(\theta, \mathbf{u}, p)$ via the adjoint variable method. Let the Lagrange function of f_0 be

$$\begin{aligned} \mathcal{L}_0(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0) &= f_0(\theta, \mathbf{u}, p) - \mathcal{L}_S(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0) \\ &= \int_D \left\{ \mu \left(\nabla \mathbf{u}^\top \right) \cdot \left(\nabla \mathbf{v}_0^\top \right) + \psi(\phi(\theta)) \mathbf{u} \cdot \mathbf{v}_0 - p \nabla \cdot \mathbf{v}_0 \right. \\ &\quad \left. - \mathbf{b} \cdot (\mathbf{v}_0 + \mathbf{u}) - q_0 \nabla \cdot \mathbf{u} \right\} dx \\ &\quad - \int_{\partial D} \{ (\mathbf{u} - \mathbf{u}_D) \cdot (\mu \partial_\nu \mathbf{v}_0 - q_0 \mathbf{v}) + (\mathbf{v}_0 - \mathbf{u}_D) \cdot (\mu \partial_\nu \mathbf{u} - p \mathbf{v}) \} d\gamma. \end{aligned} \quad (8.10.11)$$

Comparing with Eq. (8.9.8) defining the Lagrange function with respect to mean compliance of linear elastic body, here a negative sign was put on \mathcal{L}_S . This change is so that a self-adjoint relationship can be obtained later. Although in the mean compliance minimization problem of a linear elastic body, the minimization of the displacement was the aim, in a Stokes flow field mean flow resistance minimization problem, the maximization of flow is aimed for. Hence, this type of difference arose. The Fréchet derivative of \mathcal{L}_0 with respect to an arbitrary variation $(\vartheta, \hat{\mathbf{u}}, \hat{p}, \hat{\mathbf{v}}_0, \hat{q}_0) \in X \times (U \times P)^2$ of $(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0)$ is

$$\begin{aligned} \mathcal{L}'_0(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0) &[\vartheta, \hat{\mathbf{u}}, \hat{p}, \hat{\mathbf{v}}_0, \hat{q}_0] \\ &= \mathcal{L}_{0\theta}(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0) [\vartheta] + \mathcal{L}_{0\mathbf{u}p}(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0) [\hat{\mathbf{u}}, \hat{p}] \\ &\quad + \mathcal{L}_{0\mathbf{v}_0q_0}(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0) [\hat{\mathbf{v}}_0, \hat{q}_0]. \end{aligned} \quad (8.10.12)$$

Each term is considered below.

The third term on the right-hand side of Eq. (8.10.12) becomes

$$\begin{aligned}\mathcal{L}_{0v_0q_0}(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0) [\hat{\mathbf{v}}_0, \hat{q}_0] &= \mathcal{L}_{Sv_0q_0}(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0) [\hat{\mathbf{v}}_0, \hat{q}_0] \\ &= -\mathcal{L}_S(\theta, \mathbf{u}, p, \hat{\mathbf{v}}_0, \hat{q}_0).\end{aligned}\quad (8.10.13)$$

Equation (8.10.13) is a Lagrange function of the state determination problem (Problem 8.10.2). Hence, if (\mathbf{u}, p) is the weak solution of the state determination problem, the third term on the right-hand side of Eq. (8.10.12) vanishes.

Moreover, the second term on the right-hand side of Eq. (8.10.12) becomes

$$\begin{aligned}\mathcal{L}_{0up}(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0) [\hat{\mathbf{u}}, \hat{p}] &= \int_D \left\{ \mu \left(\nabla \hat{\mathbf{u}}^\top \right) \cdot \left(\nabla \mathbf{v}_0^\top \right) + \psi(\phi(\theta)) \hat{\mathbf{u}} \cdot \mathbf{v}_0 - \hat{p} \nabla \cdot \mathbf{v}_0 \right. \\ &\quad \left. - \mathbf{b} \cdot \hat{\mathbf{u}} - q_0 \nabla \cdot \hat{\mathbf{u}} \right\} dx \\ &\quad - \int_{\partial D} \left\{ \hat{\mathbf{u}} \cdot (\mu \partial_\nu \mathbf{v}_0 - q_0 \mathbf{v}) + (\mathbf{v}_0 - \mathbf{u}_D) \cdot (\mu \partial_\nu \hat{\mathbf{u}} - \hat{p} \mathbf{v}) \right\} d\gamma \\ &= -\mathcal{L}_S(\theta, \mathbf{v}_0, q_0, \hat{\mathbf{u}}, \hat{p})\end{aligned}\quad (8.10.14)$$

with respect to an arbitrary variation $(\hat{\mathbf{u}}, \hat{p}) \in U \times P$ of (\mathbf{u}, p) . Hence, when the self-adjoint relationship

$$(\mathbf{u}, p) = (\mathbf{v}_0, q_0) \quad (8.10.15)$$

holds, the second term on the right-hand side of Eq. (8.10.12) becomes zero.

Furthermore, the first term on the right-hand side of Eq. (8.10.12) becomes

$$\mathcal{L}_{0\theta}(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0) [\vartheta] = \int_D \psi'(\phi(\theta)) \phi'(\theta) \mathbf{u} \cdot \mathbf{v}_0 \vartheta \, dx. \quad (8.10.16)$$

Therefore, suppose (\mathbf{u}, p) is a weak solution of Problem 8.10.2 and that the self-adjoint relationship Eq. (8.10.15) holds. If $f_0(\theta, \mathbf{u}, p)$ in this case is written as $\tilde{f}_0(\theta)$, the equation

$$\tilde{f}'_0(\theta) [\vartheta] = \mathcal{L}_{0\theta}(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0) [\vartheta] = \langle g_0, \vartheta \rangle \quad (8.10.17)$$

holds, where

$$g_0 = \psi' \phi' \mathbf{u} \cdot \mathbf{u}. \quad (8.10.18)$$

When using $\psi(\phi)$ of Eq. (8.10.3), we get

$$\psi'(\phi) = -\psi_1 \frac{\alpha(1+\alpha)}{(\phi+\alpha)^2}. \quad (8.10.19)$$

On the other hand, with respect to $f_1(\theta)$,

$$f'_1(\theta)[\vartheta] = \int_D \phi' \vartheta \, dx \quad (8.10.20)$$

holds with respect to an arbitrary $\vartheta \in X$.

Based on the results above, the function space containing g_0 of Eq. (8.10.18) is included in $W^{1,q_R}(D; \mathbb{R}) \subset X'$ which is smoother than the result of Theorem 8.5.2. From the fact that $W^{1,q_R}(D; \mathbb{R}) \subset C^0(D; \mathbb{R})$, it is thought that even without applying the H^1 gradient method, a numerically unstable phenomenon will not be generated. However, in order for the search vector ϑ_g to guarantee $C^{0,1}$ class, the H^1 gradient method is required.

8.10.4 Second-Order θ -Derivatives of Cost Functions

Furthermore, the second-order θ -derivatives of the cost functions of the mean flow resistance f_0 and constraint function f_1 with respect to the domain measure of flow field can be obtained. Based on the procedures looked at in Sect. 8.5.2, let us also look here at obtaining the second-order θ derivatives of f_0 and f_1 .

Firstly, let us think about the second-order θ -derivative of f_0 . With respect to Hypothesis 8.5.4 (1), \mathbf{b} is assumed not to be a function of θ . The relationship corresponding to Hypothesis 8.5.4 (2) is satisfied here.

The Lagrange function \mathcal{L}_0 of f_0 is defined in Eq. (8.10.11). Viewing (θ, \mathbf{u}, p) as a design variable, its admissible set and admissible direction is set as

$$S = \{(\theta, \mathbf{u}, p) \in \mathcal{D} \times \mathcal{S} \times \mathcal{Q} \mid \mathcal{L}_S(\theta, \mathbf{u}, p, \mathbf{v}, q) = 0 \text{ for all } (\mathbf{v}, q) \in U \times P\},$$

$$T_S(\theta, \mathbf{u}, p) = \{(\vartheta, \hat{\mathbf{v}}, \hat{\pi}) \in X \times U \times P \mid \mathcal{L}_{S\theta\mathbf{u}p}(\theta, \mathbf{u}, p, \mathbf{v}, q)[\vartheta, \hat{\mathbf{v}}, \hat{\pi}] = 0 \text{ for all } (\mathbf{v}, q) \in U \times P\}.$$

The second-order θ -derivative of \mathcal{L}_0 with respect to arbitrary variations $(\vartheta_1, \hat{\mathbf{v}}_1, \hat{\pi}_1), (\vartheta_2, \hat{\mathbf{v}}_2, \hat{\pi}_2) \in T_S(\theta, \mathbf{u}, p)$ becomes

$$\begin{aligned} & \mathcal{L}_{0(\theta, \mathbf{u}, p)(\theta, \mathbf{u}, p)}(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0)[(\vartheta_1, \hat{\mathbf{v}}_1, \hat{\pi}_1), (\vartheta_2, \hat{\mathbf{v}}_2, \hat{\pi}_2)] \\ &= \mathcal{L}_{0\theta\theta}(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0)[\vartheta_1, \vartheta_2] + \mathcal{L}_{0\theta\mathbf{u}p}(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0)[\vartheta_1, \hat{\mathbf{v}}_2, \hat{\pi}_2] \\ &+ \mathcal{L}_{0\theta\mathbf{u}p}(\theta, \mathbf{u}, \mathbf{v}_0)[\vartheta_2, \hat{\mathbf{v}}_1, \hat{\pi}_1] + \mathcal{L}_{0\mathbf{u}\mathbf{u}p}(\theta, \mathbf{u}, \mathbf{v}_0)[\hat{\mathbf{v}}_1, \hat{\pi}_1, \hat{\mathbf{v}}_2, \hat{\pi}_2]. \end{aligned} \quad (8.10.21)$$

Each term on the right-hand side of Eq. (8.10.21) becomes

$$\begin{aligned} \mathcal{L}_{0\theta\theta}(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0) [\vartheta_1, \vartheta_2] \\ = \int_D \left\{ \psi''(\phi(\theta)) (\phi'(\theta))^2 + \psi'(\phi(\theta)) \phi''(\theta) \right\} \mathbf{u} \cdot \mathbf{v}_0 \vartheta_1 \vartheta_2 \, dx, \end{aligned} \quad (8.10.22)$$

$$\mathcal{L}_{0\theta\mathbf{u}p}(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0) [\vartheta_1, \hat{\mathbf{v}}_2, \hat{\pi}_2] = \int_D \psi'(\phi(\theta)) \phi'(\theta) \hat{\mathbf{v}}_2 \cdot \mathbf{v}_0 \vartheta_1 \, dx, \quad (8.10.23)$$

$$\mathcal{L}_{0\theta\mathbf{u}p}(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0) [\vartheta_2, \hat{\mathbf{v}}_1, \hat{\pi}_1] = \int_D \psi'(\phi(\theta)) \phi'(\theta) \hat{\mathbf{v}}_1 \cdot \mathbf{v}_0 \vartheta_2 \, dx, \quad (8.10.24)$$

$$\mathcal{L}_{0\mathbf{u}p\mathbf{u}p}(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0) [\hat{\mathbf{v}}_1, \hat{\pi}_1, \hat{\mathbf{v}}_2, \hat{\pi}_2] = 0. \quad (8.10.25)$$

Here, the fact that $\mathbf{u} - \mathbf{u}_D$, $\mathbf{v}_0 - \mathbf{u}_D$, $\hat{\mathbf{v}}_1$ and $\hat{\mathbf{v}}_2$ become $\mathbf{0}_{\mathbb{R}^d}$ on ∂D was used. In this case, with respect to arbitrary variation $(\vartheta_j, \hat{\mathbf{v}}_j, \hat{\pi}_j) \in T_S(\theta, \mathbf{u}, p)$ for $j \in \{1, 2\}$, the Fréchet partial derivative of the Lagrange function \mathcal{L}_S of the state determination problem establishes the equation

$$\begin{aligned} \mathcal{L}_{S\theta\mathbf{u}p}(\theta, \mathbf{u}, p, \mathbf{v}, q) [\vartheta, \hat{\mathbf{v}}, \hat{\pi}] \\ = \int_D \left\{ -\mu (\nabla \hat{\mathbf{v}}^\top) \cdot (\nabla \mathbf{v}^\top) - \psi'(\phi(\theta)) \phi'(\theta) \mathbf{u} \cdot \mathbf{v} \vartheta \right. \\ \left. - \psi(\phi(\theta)) \hat{\mathbf{v}} \cdot \mathbf{v} + \hat{\pi} \nabla \cdot \mathbf{v} + q \nabla \cdot \hat{\mathbf{v}} \right\} dx \\ = 0 \end{aligned} \quad (8.10.26)$$

with respect to an arbitrary $(\mathbf{v}, q) \in U \times P$. Here, the fact that \mathbf{v} and $\hat{\mathbf{v}}$ are $\mathbf{0}_{\mathbb{R}^d}$ on Γ_D as well as Eq. (8.10.5) were used.

Here, the next assumption is set up. At a local minimum point of a mean flow resistance minimization problem (Problem 8.10.3), it is thought that the fluid content $\phi(\theta)$ converges to 0 and 1, and the terms introduced in order to provide flow resistance becomes sufficiently small in actual flow field of $\phi(\theta) \approx 1$. Hence, in Eq. (8.10.26), it is assumed that

$$\int_D \left\{ -\mu (\nabla \hat{\mathbf{v}}_j^\top) \cdot (\nabla \mathbf{v}^\top) + \hat{\pi}_j \nabla \cdot \mathbf{v} + q \nabla \cdot \hat{\mathbf{v}}_j \right\} dx = 0 \quad (8.10.27)$$

is established. In this case, the condition

$$\hat{\mathbf{v}}_j = -\frac{\psi'(\phi(\theta)) \phi'(\theta)}{\psi(\phi(\theta))} \vartheta_j \mathbf{u} \quad \text{in } D \quad (8.10.28)$$

can be obtained. Hence, substituting $(\hat{\mathbf{v}}_j, \hat{\pi}_j)$ into $(\hat{\mathbf{v}}_1, \hat{\pi}_1)$ in Eq. (8.10.24) and $(\hat{\mathbf{v}}_2, \hat{\pi}_2)$ in Eq. (8.10.23), then

$$\begin{aligned}\mathcal{L}_{0\theta\mathbf{u}_p}(\theta, \mathbf{u}, \mathbf{v}_0) & [\vartheta_1, \hat{\mathbf{v}}_2, \hat{\pi}_2] \\ & = \mathcal{L}_{0\theta\mathbf{u}_p}(\theta, \mathbf{u}, p, \mathbf{v}_0, q_0) [\vartheta_2, \hat{\mathbf{v}}_1, \hat{\pi}_1] \\ & = \int_D -\frac{(\psi'(\phi(\theta))\phi'(\theta))^2}{\psi(\phi(\theta))} \mathbf{u} \cdot \mathbf{v}_0 \vartheta_1 \vartheta_2 \, dx\end{aligned}\quad (8.10.29)$$

is obtained.

Summarizing the results above, by substituting Eqs. (8.10.29) and (8.10.22) into Eq. (8.10.21), the second-order θ -derivative of mean flow resistance f_0 becomes

$$\begin{aligned}h_0(\vartheta_1, \vartheta_2) & = \int_D \left\{ \psi''(\phi')^2 + \psi'\phi'' - 2\frac{(\psi'\phi')^2}{\psi} \right\} \mathbf{u} \cdot \mathbf{v}_0 \vartheta_1 \vartheta_2 \, dx \\ & = \int_D \psi_1 \beta(\alpha, \theta) \mathbf{u} \cdot \mathbf{v}_0 \vartheta_1 \vartheta_2 \, dx.\end{aligned}\quad (8.10.30)$$

In the above equation, $\psi'(\phi)$ becomes Eq. (8.10.19) and

$$\psi''(\phi) = \frac{2\alpha(1+\alpha)}{(\phi+\alpha)^3}. \quad (8.10.31)$$

When Eq. (8.1.1) is used in $\phi(\theta)$, $\phi'(\theta)$ and $\phi''(\theta)$ are given by Eqs. (8.5.7) and (8.9.27), respectively. Moreover, if $\phi(\theta)$ is given by Eq. (8.1.2), these are given by Eqs. (8.5.8) and (8.9.28), respectively. Figure 8.19 shows the graph of $\beta(\alpha, \theta)$. From $\beta(\alpha, \theta) < 0$ and the self-adjoint relationship $\mathbf{u} \cdot \mathbf{v}_0 = \mathbf{u} \cdot \mathbf{u} > 0$, it can be confirmed that $h_0(\cdot, \cdot)$ is not coercive.

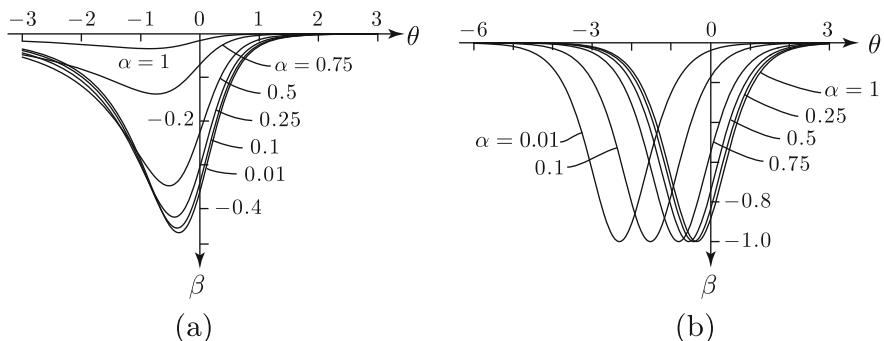


Fig. 8.19 Coefficient functions $\beta(\alpha, \theta)$ at second-order θ -derivative of mean flow resistance. **(a)** $\phi(\theta) = \tan^{-1}\theta/\pi + 1/2$. **(b)** $\phi(\theta) = (\tanh\theta + 1)/2$

On the other hand, the second-order θ -derivative of $f_1(\theta)$ becomes Eq. (8.9.26). The graph of $\phi''(\theta)$ is shown in Fig. 8.13.

In this way, the second-order θ -derivative of the cost function f_0 in mean flow resistance minimization problem is not coercive and the second-order θ -derivative of the constraint function f_1 does not become coercive either. Hence, if the Newton method (Problem 8.6.6) is to be used with respect to a mean flow resistance minimization problem, there is a need to use an appropriate bilinear form $a_X(\vartheta_{gi}, \psi)$ to capture coerciveness.

8.10.5 Second-Order θ -Derivative of Cost Function Using Lagrange Multiplier Method

When the Lagrange multiplier method is used to obtain the second-order θ -derivative of the mean flow resistance f_0 , it becomes as follows. Using the same discussion as Sect. 7.5.4, we fix ϑ_1 and define the Lagrange function for $\tilde{f}'_0(\theta)[\vartheta_1] = \langle g_0, \vartheta_1 \rangle$ in Eq. (8.10.17) by

$$\mathcal{L}_{10}(\theta, \mathbf{u}, p, \mathbf{w}_0, r_0) = \langle g_0, \vartheta_1 \rangle - \mathcal{L}_S(\theta, \mathbf{u}, p, \mathbf{w}_0, r_0), \quad (8.10.32)$$

where \mathcal{L}_S is given by Eq. (8.10.8), and $(\mathbf{w}_0, r_0) \in U \times P$ is the adjoint variable provided for (\mathbf{u}, p) in \mathbf{g}_0 .

With respect to arbitrary variations $(\vartheta_2, \hat{\mathbf{u}}, \hat{p}, \hat{\mathbf{w}}_0, \hat{r}_0) \in X \times (U \times P)^2$ of $(\theta, \mathbf{u}, p, \mathbf{w}_0, r_0)$, the Fréchet derivative of \mathcal{L}_{10} is written as

$$\begin{aligned} \mathcal{L}'_{10}(\theta, \mathbf{u}, p, \mathbf{w}_0, r_0) & [\vartheta_2, \hat{\mathbf{u}}, \hat{p}, \hat{\mathbf{w}}_0, \hat{r}_0] \\ &= \mathcal{L}_{10\theta}(\theta, \mathbf{u}, p, \mathbf{w}_0, r_0) [\vartheta_2] + \mathcal{L}_{10\mathbf{u}p}(\theta, \mathbf{u}, p, \mathbf{w}_0, r_0) [\hat{\mathbf{u}}, \hat{p}] \\ &+ \mathcal{L}_{10\mathbf{w}_0r_0}(\theta, \mathbf{u}, p, \mathbf{w}_0, r_0) [\hat{\mathbf{w}}_0, \hat{r}_0]. \end{aligned} \quad (8.10.33)$$

The third term on the right-hand side of Eq. (8.10.33) vanishes if (\mathbf{u}, p) is the solution of the state determination problem.

The second term on the right-hand side of Eq. (8.10.33) is

$$\begin{aligned} & \mathcal{L}_{10\mathbf{u}p}(\theta, \mathbf{u}, p, \mathbf{w}_0, r_0) [\hat{\mathbf{u}}, \hat{p}] \\ &= \int_D \left\{ 2\psi' \phi' \mathbf{u} \cdot \hat{\mathbf{u}} \vartheta_1 + \mu \left(\nabla \mathbf{w}_0^\top \right) \cdot \left(\nabla \hat{\mathbf{u}}^\top \right) + \psi(\phi(\theta)) \mathbf{w}_0 \cdot \hat{\mathbf{u}} \right. \\ & \quad \left. - \hat{p} \nabla \cdot \mathbf{w}_0 - r_0 \nabla \cdot \hat{\mathbf{u}} \right\} dx. \end{aligned} \quad (8.10.34)$$

Here, the condition that Eq. (8.10.34) is zero for arbitrary $(\hat{\mathbf{u}}, \hat{p}) \in U \times P$ is equivalent to setting (\mathbf{w}_0, r_0) to be the solution of the following adjoint problem.

Problem 8.10.4 (Adjoint Problem of (\mathbf{w}_0, r_0) with Respect to $\langle g_0, \vartheta_1 \rangle$) Under the assumption of Problem 8.10.2, let $\vartheta_1 \in X$ be given. Find $(\mathbf{w}_0, r_0) = (\mathbf{w}_0(\vartheta_1), r_0(\vartheta_1)) \in U \times P$ satisfying

$$\begin{aligned} -\nabla^\top (\mu \nabla \mathbf{w}_0^\top) + \psi \mathbf{w}_0^\top + \nabla^\top r_0 &= -2\psi' \phi' \mathbf{v}_0^\top \vartheta_1 \quad \text{in } D, \\ \nabla \cdot \mathbf{w}_0 &= 0 \quad \text{in } D, \\ \mathbf{w}_0 &= \mathbf{0}_{\mathbb{R}^d} \quad \text{on } \partial D, \\ \int_D r_0 \, dx &= 0. \end{aligned}$$

□

Finally, the first term on the right-hand side of Eq. (8.10.33) becomes

$$\begin{aligned} \mathcal{L}_{10\theta}(\theta, \mathbf{u}, p, \mathbf{w}_0, r_0)[\vartheta_2] \\ = \int_D \left\{ \left(\psi''(\phi')^2 + \psi' \phi'' \right) \mathbf{u} \cdot \mathbf{u} \vartheta_1 + \psi' \phi' \mathbf{u} \cdot \mathbf{w}_0(\vartheta_1) \right\} \vartheta_2 \, dx. \end{aligned}$$

Here, (\mathbf{u}, p) and $(\mathbf{w}_0(\vartheta_1), r_0(\vartheta_1))$ are assumed to be the weak solutions of Problems 8.10.2 and 8.10.4, respectively. If we denote $f_i(\theta, \mathbf{u}, p)$ here by $\tilde{f}_i(\theta)$, we have the relation:

$$\begin{aligned} \mathcal{L}_{10\theta}(\theta, \mathbf{u}, p, \mathbf{w}_0(\vartheta_1), r_0)[\vartheta_2] &= \tilde{f}_0''(\theta)[\vartheta_1, \vartheta_2] \\ &= \langle g_{H0}(\theta, \vartheta_1), \vartheta_2 \rangle, \end{aligned} \quad (8.10.35)$$

where the Hesse gradient g_{H0} of the mean flow resistance is given by

$$g_{H0}(\theta, \vartheta_1) = \left(\psi''(\phi')^2 + \psi' \phi'' \right) \mathbf{u} \cdot \mathbf{u} \vartheta_1 + \psi' \phi' \mathbf{u} \cdot \mathbf{w}_0(\vartheta_1). \quad (8.10.36)$$

If the same relation with Eq. (8.10.27) is satisfied in Problem 8.10.4,

$$\mathbf{w}_0(\vartheta_1) = -\frac{\psi' \phi'}{\psi} \mathbf{v}_0 \vartheta_1 \quad (8.10.37)$$

holds. Substituting Eq. (8.10.37) into Eq. (8.10.36), it can be confirmed that Eq. (8.10.35) accords with Eq. (8.10.30).

8.10.6 Numerical Example

The results of mean flow resistance minimization for a two-dimensional Stokes flow field around an isolated object are shown in Figs. 8.20, 8.21, 8.22, 8.23. The boundary condition of a state determination problem is assumed on the outer

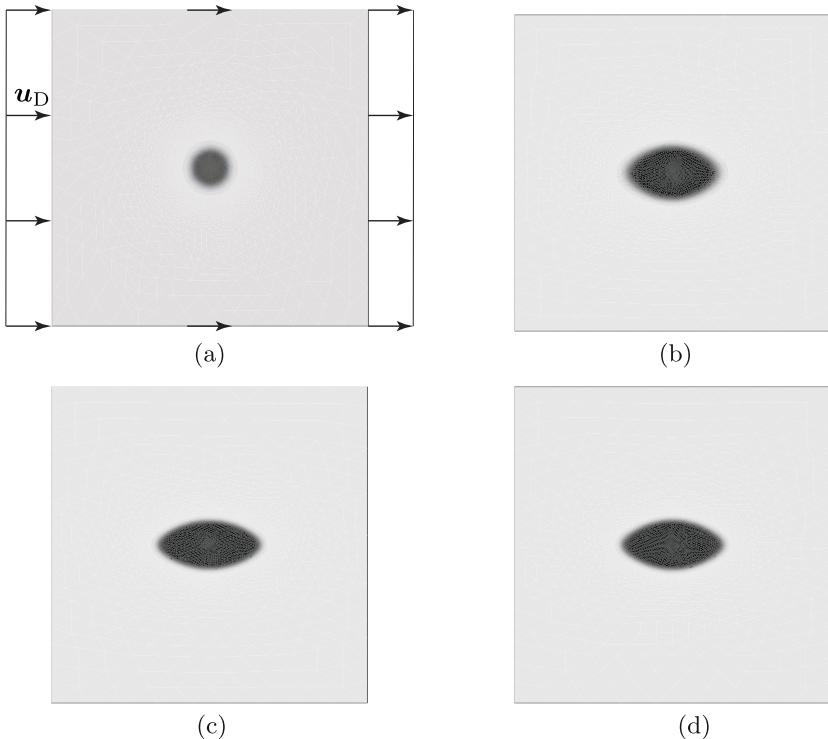


Fig. 8.20 Numerical example of mean flow resistance minimization: densities (black and white are inverted). (a) Initial density and boundary condition. (b) H^1 gradient method ($k = 100$). (c) H^1 Newton method ($k = 100$). (d) H^1 Newton method (Hesse gradient, $k = 100$)

boundary with a uniform flow field in the horizontal direction as shown in Fig. 8.20a. With respect to the initial θ , a two-dimensional Gaussian distribution was assumed in order for it to be an element of the admissible set \mathcal{D} . Also in this example, a domain in which the density is constrained was not set. The programs were written using the programming language FreeFEM (<https://freefem.org/>) [66] using the finite element method. In the finite element analyses of the Stokes problem, triangular elements of the second order with respect to the velocity and of the first order with respect to the pressure were used. In the case using the H^1 Newton method, the routine of the H^1 Newton method was started at $k_N = 20$. For further details, we recommend the readers to also examine the exact codes in the programs.²

Figure 8.20b–d show the densities obtained by the three methods (H^1 gradient method using $g_{\mathcal{L}} = g_0 + \lambda_1 g_1$, H^1 Newton method using $h_{\mathcal{L}} = h_0 + \lambda_1 h_1$ and $g_{\mathcal{L}}$, and the H^1 Newton method using $g_{H0}, h_1, g_{\mathcal{L}}$). In Fig. 8.21, the streamlines

²See Electronic supplementary material.

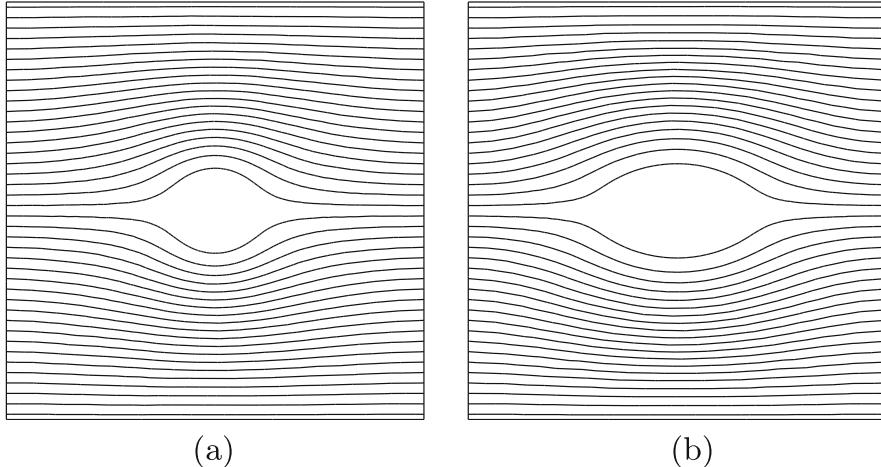


Fig. 8.21 Numerical example of mean flow resistance minimization problem: streamlines **(a)** Initial density. **(b)** H^1 Newton method

at the initial density and at the optimal density obtained from the H^1 Newton method are shown, respectively. The streamlines are defined as the contour lines of the flow function $\psi : \Omega(\phi) \rightarrow \mathbb{R}$ such that the flow speed \mathbf{u} is given by $(\partial\psi/\partial x_2, -\partial\psi/\partial x_1)^\top$.

The graphs in Fig. 8.22 illustrate the histories of the cost functions as well as the gradients and Hessians of the objective function f_0 on the search path with respect to the iteration number k and the search distance $\sum_{i=0}^{k-1} \|\vartheta_{g(i)}\|_X$ on X . In this figure, $f_{0\text{init}}$ denotes the value of f_0 at the initial density. Also, the value of c_1 is taken to be the domain integral of the initial density (volume). The gradient of f_0 on the search path was calculated using the Lagrange function $\mathcal{L} = \mathcal{L}_0 + \lambda_1 f_1$ as $\langle g_0 + \lambda_1 g_1, \vartheta_{g(k)} \rangle / \|\vartheta_{g(k)}\|_X$. On the other hand, the Hessian of f_0 on the search path was computed via $h_{\mathcal{L}}[\vartheta_{g(k)}, \vartheta_{g(k)}] / \|\vartheta_{g(k)}\|_X^2$. In the case of the Newton method using the Hesse gradient, the ratio $(\langle g_{00}, \vartheta_{g(k)} \rangle + \lambda_1 h_1[\vartheta_{g(k)}, \psi]) / \|\vartheta_{g(k)}\|_X^2$ was used to calculate the Hessian. The norm $\|\vartheta_{g(i)}\|_X$ of the i -th search vector is defined by Eq. (8.9.35). The computational times until $k = 100$ by PC were 10.839, 14.763 and 17.046 sec by the H^1 gradient method, the H^1 Newton method and the H^1 Newton method using the Hesse gradient, respectively.

Looking at the graphs in Fig. 8.22c, it also seems that the convergence speed with respect to the iteration number k is faster when the H^1 Newton method is applied than when the H^1 gradient method is used. However, this increase in convergence speed might have been due to the fact that c_{D1} and c_{D0} in Eq. (8.6.7) are set with smaller values that had made the step sizes larger. Meanwhile, we noticed that the search paths for the three methods are the same as evident from the three coinciding graphs plotted in Fig. 8.22d. Moreover, from Fig. 8.22d,f, it can be observed that the

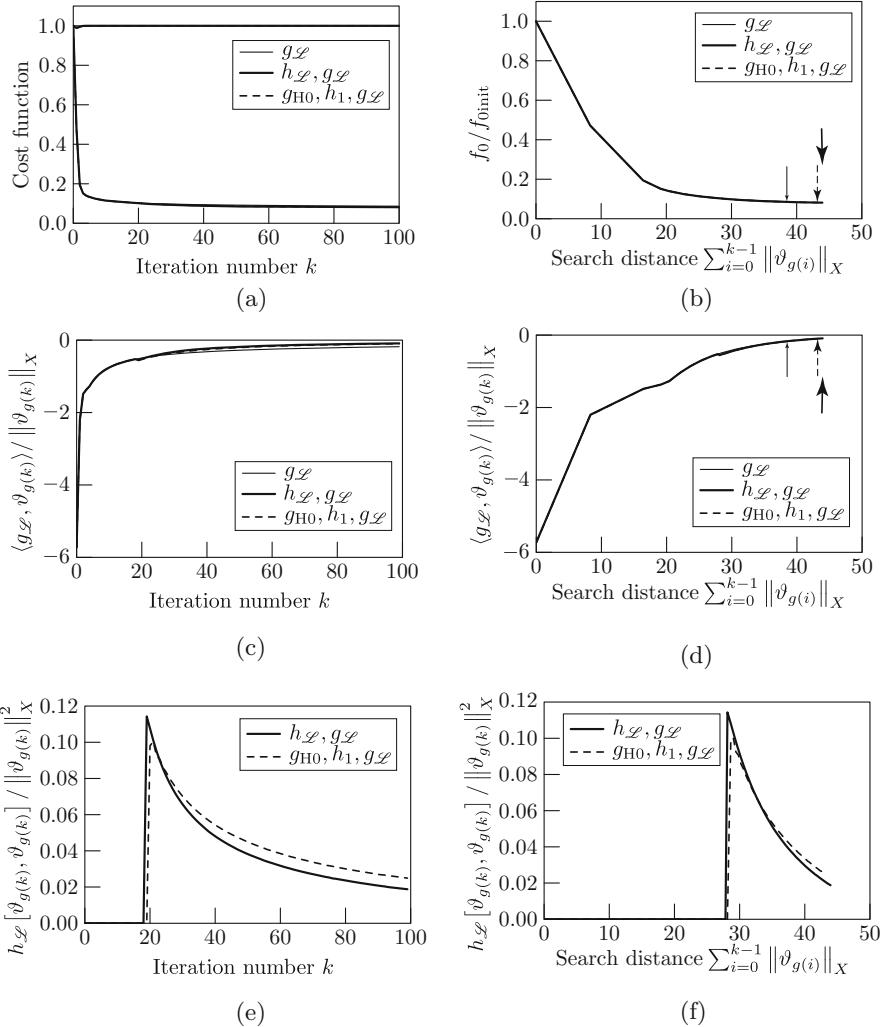


Fig. 8.22 Numerical example of mean flow resistance minimization: cost functions and gradients and Hessians of f_0 on the search path ($g_{\mathcal{L}}$: H^1 gradient method, $h_{\mathcal{L}}, g_{\mathcal{L}}$: H^1 Newton method, $g_{H0}, h_1, g_{\mathcal{L}}$: H^1 Newton method using Hesse gradient). (a) Cost functions. (b) Cost functions (search distance). (c) Gradient of f_0 on the search path. (d) Gradient of f_0 on the search path (search distance). (e) Hessian of f_0 on the search path. (f) Hessian of f_0 on the search path (search distance)

point of convergence is a local minimum and that both the first derivative and the Hessian eventually diminish to zero. The reason behind these observed behaviors of the methods is the same as the ones given at the end of Sect. 8.9.6. Moreover, in the case of the Stokes flow field, although $h_0(\cdot, \cdot)$ itself is not coercive as shown in

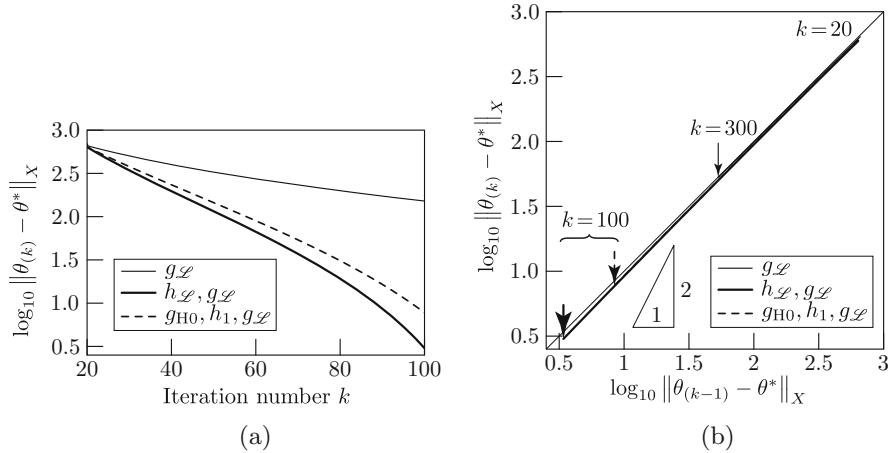


Fig. 8.23 Numerical example of mean flow resistance minimization problem: distance $\|\theta^{(k)} - \theta^*\|_X$ from an approximate minimum point θ^* (g_L : the gradient method, h_L, g_L : the Newton method, g_{H0}, h_1, g_L : the Newton method using the Hesse gradient). (a) Iteration history. (b) $(k-1)$ -th vs. k -th plot

Sect. 8.10.4, it can be confirmed that the point of convergence is a minimum point, since the Hessian of f_0 on the search path is positive valued.

In addition, Fig. 8.23a shows the graphs of the distance $\|\theta^{(k)} - \theta^*\|_X$ from an approximate minimum point θ^* obtained by the three methods with respect to the iteration number k . The approximate minimum point θ^* was substituted with the numerical solution of θ when the iteration time is taken larger than the given value in the H^1 Newton method. From this figure, it can be confirmed that the convergence orders for the results via the H^1 Newton methods are higher compared to that of the first order. However, from Fig. 8.23b, plotting the k -th distance $\|\theta^{(k)} - \theta^*\|_X$ with respect to the $(k-1)$ -th distance, the convergence order of the H^1 Newton method is less than the second order but is greater than the first order. The reason behind this result is considered the same as that stated at the end of Sect. 8.9.6.

8.11 Summary

In Chap. 8, the problem for seeking optimal hole positions with respect to a domain, in which a boundary value problem of partial differential equation is defined, is constructed as topology optimization problem of θ -type and its solution was looked at in detail. The key points are as below:

- (1) If the characteristic function of a domain is chosen as a design variable, it is known that regularity is insufficient and an optimization problem cannot be constructed (beginning of Chap. 8).

- (2) If a density is chosen as the design variable, a topology optimization problem can be constructed (beginning of Chap. 8). However, the set of functions whose range is limited to $[0, 1]$ does not become a linear space. Hence, by choosing a function $\theta \in X = H^1(D; \mathbb{R})$ whose range is not constrained as the design variable and providing the density as a sigmoid function of θ , a topology optimization problem of θ -type can be constructed based on the framework of the abstract optimal design problem shown in Chap. 7 (Sect. 8.1).
- (3) When a Poisson problem is chosen as a state determination problem (Sect. 8.2), a topology optimization problem of θ -type will be constructed as Problem 8.3.2 (Sect. 8.3).
- (4) The θ -derivatives of cost functions can be obtained by the Lagrange multiplier method (Sect. 8.5.1). However, such θ -derivatives are not necessarily going to be in X (Remark 8.5.3). Moreover, the second-order θ -derivatives of cost functions can be sought by substituting in the θ -derivative of the solution of a state determination problem into the second-order θ -derivative of the Lagrange function (Sect. 8.5.2).
- (5) The descent vectors of cost functions can be obtained using the θ -derivatives of cost functions by a gradient method (H^1 gradient method) on $X = H^1(D; \mathbb{R})$ (Sect. 8.6.1). The solution of the H^1 gradient method is included in the admissible set apart from the singular points (Theorem 8.6.5). Furthermore, if the second-order θ -derivatives of cost functions can be calculated, the descent vectors of the cost functions can be sought via the H^1 Newton method (Sect. 8.6.2).
- (6) The solution to the topology optimization problem of θ -type can be constructed using the same framework as the gradient method with respect to a constrained problem and the Newton method with respect to a constrained problem shown in Chap. 3 (Sects. 8.7.1 and 8.7.2).
- (7) When the numerical solutions of a state determination problem, adjoint problem, and H^1 gradient method are to be sought via the finite element method, and a first-order finite element is used to seek the search vector ϑ_g , the error of the finite element solution reduces linearly with respect to the maximum diameter of the finite element (Theorem 8.8.5).
- (8) When a linear elastic problem is taken to be a state determination problem, and a mean compliance and a function for domain measure are chosen as object and constraint cost functions, the θ -derivatives and second-order θ -derivatives of the cost functions can be obtained (Sect. 8.9.3).
- (9) If the Stokes problem is taken to be a state determination problem, and a mean flow resistance minimization problem under a domain measure constraint is considered, the θ -derivatives and second-order θ -derivatives of the cost functions can be obtained (Sect. 8.10.3).

8.12 Practice Problems

8.1 When a θ -type Poisson problem (Problem 8.2.3) is made into a state determination problem, what would be the cost function such that the self-adjoint relationship holds? Moreover, show its θ -derivative.

8.2 Change the extended Poisson problem defined in Chap. 5 (Problem 5.1.3) to θ -type, and formulate a topology optimization problem of θ -type using the extended Poisson problem as a state determination problem, the object cost function such that the self-adjoint relationship holds, and the constraint cost function with respect to the domain measure. Moreover, show the KKT conditions with respect to that problem.

8.3 When many cost functions are defined and the maximum values of them have to be minimized, the β method is known as a way to construct an optimal design problem [164]. If the topology optimization problem of θ -type is rewritten with the β method, we obtain the following:

Problem 8.12.1 (The β Method) Let \mathcal{D} and \mathcal{S} be Eqs. (8.1.4) and (8.2.2), respectively, and $f_1, \dots, f_m : X \times \mathcal{S} \rightarrow \mathbb{R}$ be given as Eq. (8.3.1). Moreover, let $\beta \in \mathbb{R}$. In this case, obtain θ which satisfies:

$$\min_{(\theta, u - u_D) \in \mathcal{D} \times \mathcal{S}} \{ \beta \mid f_1(\theta, u) \leq \beta, \dots, f_m(\theta, u) \leq \beta, \text{ Problem 8.2.3} \}. \quad \square$$

Show the KKT conditions with respect to this problem. Moreover, show the method to determine the Lagrange multipliers when solving this problem using the H^1 gradient method (Sect. 8.7.1) with respect to a constrained problem.

(Supplement) The reason that the β method is preferred is shown as follows. Even if there are many cost functions, the Lagrange multipliers with respect to cost functions for which inequality constraints are not active are zero. Hence, there are no reasons to seek their θ -derivatives.

8.4 In Practice 1.2, the gradient \mathbf{g}_0 of mean compliance f_0 with respect to variation of cross-sectional area was sought by using the gradient of maximization problem of the potential energy with respect to variation of \mathbf{a} and the minimum condition of the potential energy with respect to variation of \mathbf{u} . With respect to the mean compliance minimization problem (Problem 8.9.3) of a $d \in \{2, 3\}$ -dimensional linear elastic body, think about the problem using

$$\begin{aligned} \pi(\theta, \mathbf{u}) = & \int_D \left(\frac{1}{2} \phi^\alpha(\theta) \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{u}) - \mathbf{b}(\theta) \cdot \mathbf{u} \right) dx \\ & - \int_{\Gamma_N} \mathbf{p}_N \cdot \mathbf{u} d\gamma - \int_{\Gamma_D} (\mathbf{u} - \mathbf{u}_D) \cdot (\phi^\alpha(\theta) \mathbf{S}(\mathbf{v}_0) \mathbf{v}) d\gamma \end{aligned}$$

as the potential energy and seeking (θ, \mathbf{u}) that satisfies

$$\max_{\theta \in \mathcal{D}} \min_{\mathbf{u} \in U} \pi(\theta, \mathbf{u}).$$

In this case, show that the θ gradient when using \mathbf{u} satisfying $\min_{\mathbf{u} \in U} \pi$ is the same as half of g_0 in Eq. (8.9.14).

8.5 In Practice 1.8, the gradient g_0 of mean flow resistance f_0 with respect to variation of cross-sectional area was obtained using the gradient of minimization problem of a formal potential energy for a dissipative system with respect to variation of \mathbf{a} and maximum condition of the formal potential energy with respect to variation of \mathbf{p} . With respect to the mean flow resistance minimization problem (Problem 8.10.3) of $d \in \{2, 3\}$ -dimensional Stokes flow field,

$$\begin{aligned} \pi(\theta, \mathbf{u}, p) = \int_D \left\{ \frac{1}{2} \mu (\nabla \mathbf{u}^\top) \cdot (\nabla \mathbf{u}^\top) + \frac{1}{2} \psi(\phi(\theta)) \mathbf{u} \cdot \mathbf{u} - p \nabla \cdot \mathbf{u} \right. \\ \left. - \mathbf{b} \cdot \mathbf{u} \right\} dx - \int_{\partial D} (\mathbf{u} - \mathbf{u}_D) \cdot (\mu \partial_\nu \mathbf{u} - p \mathbf{v}) d\gamma \end{aligned}$$

is used to seek (θ, \mathbf{u}, p) which satisfies

$$\min_{\theta \in \mathcal{D}} \min_{\mathbf{u} \in U} \max_{p \in P} \pi(\theta, \mathbf{u}, p).$$

In this case, show that the θ gradient of π when using (\mathbf{u}, p) satisfying $\min_{\mathbf{u} \in U} \max_{p \in P} \pi$ is the same as half of g_0 in Eq. (8.10.17).

Chapter 9

Shape Optimization Problems of Domain Variation Type



In Chap. 8 we looked at problems for obtaining the optimal topologies of continua with the densities of continua set to be the design variable. In this chapter, we shall look at the type of shape optimization problems in which the boundary of a continuum varies. The key theory of numerical solution shown in this chapter is published in the paper [9]. In this book, we shall look at the theory used there by comparing it to the contents shown in Chaps. 1 to 7.

First, let us take an abridged look at the history of research relating to a shape optimization problem of domain variation type. This type of shape optimization problem is also referred to as a domain optimization problem and has been studied since the early twentieth century. For example, among the vast works of Hadamard, there is a description relating to a problem seeking the boundary shape of a thin membrane such that the fundamental vibration frequency is maximized. In this description, a notion equivalent to a Fréchet derivative of the fundamental frequency when a boundary is moved in the outward normal direction is presented [60, 154]. Even after that, Fréchet derivatives with respect to shape variations of domain variation type have been referred to as shape derivatives, and many researchers have announced research results relating to it.¹ To add background to this research, there are works relating to optimal control theory assuming a function as a control variable by mathematicians lead by Lions [109].

In this way, theories relating to the calculation methods of shape derivatives have been developed consistently, but research relating to moving the shapes using shape derivatives has not always obtained favorable results. In reality, it is known that if the node coordinates on a boundary of a finite element model are chosen to

Electronic Supplementary Material The online version of this article (https://doi.org/10.1007/978-981-15-7618-8_9) contains supplementary material, which is available to authorized users.

¹For example, refer to [18–20, 34–39, 43, 44, 62–65, 68, 119, 124, 132–135, 152, 154, 185, 186].

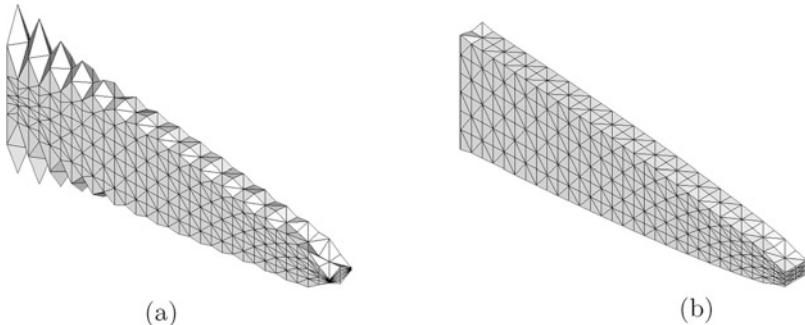


Fig. 9.1 Numerical examples with respect to the shape optimization problem of a linear elastic body (provided by Quint Corporation). (a) Rippling shape. (b) Optimal shape by H^1 gradient method

be the design variable, and the Fréchet derivatives with respect to the variation of the design variable are evaluated in order to move the nodes, a numerically unstable phenomenon in which the boundary becomes rippled such as shown in Fig. 9.1a appears [73]. Figure 9.1a shows the result of a numerical analysis with respect to a mean compliance minimization problem (Problem 9.12.2) of a three-dimensional linear elastic body. The boundary condition in the state determination problem constrains the displacement on the back edge, while a uniform downward facing nodal force (external force) on the horizontal central line of the front edge was assumed. The boundary condition in the shape variation problem restrains the variation in the normal direction on the front/back and left/right edges, and the variation on the horizontal central line on the front/back edge. Numerical analysis of a state determination problem uses the first-order finite elements. The calculation method of shape derivatives uses the formula of boundary integration form as shown later.

In order to avoid rippling boundaries such as in this case, there is a method to define the boundary shape as a B-spline curve, Bezier curve, etc. and choose its control variables as the design variables [30, 31]. There is also a method for giving the shape variation as the linear sum of the basic deformation modes and choosing the undetermined multipliers in this case as the design variables (basis vector method) [23, 65, 136, 166, 167]. All these methods have been highlighted and used in actual optimal designs. However, all the methods used derivatives with respect to parametric design variables, such as those explained in the Preface and differ from original shape derivatives.

In this chapter, we will look at a method for evaluating the shape derivatives of cost functions after having constructed a shape optimization problem of a domain variation type in which a function expressing the domain variation defined on an appropriate function space is set to be the design variable based on the framework of the abstract optimal design problem shown in Chap. 7. As a result, the shape gradient does not have enough regularity to create the following domain. This is

thought to be one of the factors generating numerical instability. In this situation, even if such a shape derivative is used, if an appropriate gradient method is used, there is the possibility that a shape optimization problem can be solved without facing numerical instability. In this chapter, this method will be the focus of our discussion.

Figure 9.1b shows the results obtained via the algorithm shown in Sect. 9.10. Boundary conditions and calculation method of the shape derivative are the same as in Fig. 9.1a. Numerical analysis of the state determination problem used second-order tetrahedral finite elements. Moreover, in numerical analysis of the H^1 gradient method using the Robin condition shown later, the first-order tetrahedral finite elements were used. Moreover, the validity relating to the selection of a finite element such as this is shown in Sect. 9.11.

The fundamental idea relating the gradient method on a function space was presented by Cea [34]. A primitive form of the gradient method can be found in Pironneau's monograph [134, p. 48 (17)]. In addition, a method called asymptotical regularization was proposed by Tautenhahn [163]. In contrast, in the 1990s, the author [8] proposed a gradient method on a function space which was referred to as the traction method, based on an engineering principle. After that, a generalization of the traction method was also introduced [14]. Furthermore, these methods have been applied in various engineering problems.² Moreover, the interpretation of the traction method in mathematics was also attempted in an existing report [80]. Here the domain mapping was assumed to be an element of the set of all continuous functions of some class, and the traction method was justified using the Gâteaux derivative of a cost function with respect to the variation of the domain mapping. In this chapter, a gradient method uses the Fréchet derivative of a cost function by defining the variation of the domain mapping in an appropriate Hilbert space. Based on this gradient method, it is apparent that the traction method was indeed a concrete example of that computational procedure.

Furthermore, a different method for constructing a shape optimization problem of domain variation type is proposed. As thought by Hadamard [60], since the next boundary shape can be determined by moving the boundary to the normal direction, one method is choosing the function that represents the amount of movement in the normal direction defined on the boundary as the design variable [118]. This method also uses the gradient method with the functionality of keeping the regularity equivalent to the gradient method shown in this chapter. However, if a finite element method is used for numerical analysis of a state determination problem, after the boundary has been moved by the gradient method, we have to consider a method for moving the finite element mesh within the domain along with the new boundary. In addition, methods using level-set functions for design variables are also being researched [4, 169, 180]. In these methods, a level-set function which is a continuous function with scalar value defined on a fixed domain is used to define the boundary with a set of points in the domain where its value is zero. Using these methods, the

²See, for example, Refs. [10, 13, 15–17, 69–72, 74–77, 81–93, 141–149, 173–178].

topology of the domain can easily be changed through joining the holes together by varying the level-set function. However, since the level-set function is defined using Euler notation (see after Definition 9.1.3), a wider domain is required than the actual domain. Moreover, in order to extract a numerical model from the level set of zero, some processes are required. Furthermore, stronger conditions with respect to the regularity of the solution for the state determination problem are required for the aforementioned two methods than for the method shown in this chapter. The reason for this is that when calculating the Fréchet derivatives of cost functions, only the formula of boundary integral type can be applicable.

This chapter is structured as follows. In Sects. 9.1 to 9.4, the definitions and formulae relating to functions and functionals defined on a moving domain are summarized. In Sect. 9.1, the definitions of admissible set of a design variable (function representing domain variation) and shape derivatives of functions and functionals are shown. There, attention will be given to the fact that there are two methods of defining the derivatives of functions defined on a moving domain with respect to domain variation. In this book, we shall refer to these notions of derivatives as “shape derivative of a function” and “partial shape derivative of a function”. Using these definitions, the formulae for shape derivatives relating to the Jacobi matrix of the domain mapping will be obtained in Sect. 9.2. Using the formulae, in Sect. 9.3, the propositions relating to shape derivatives of functions and functionals are shown. Here also, we will focus on the fact that the formulae using the shape derivatives of functions and partial shape derivatives of functions can be obtained. Section 9.4 defines several rules for variations of functions with respect to domain variation using the shape derivative of a function and the partial shape derivative of a function.

In Sects. 9.5 to 9.8, we will consider a shape optimization problem when a Poisson problem is chosen to be the state determination problem and present the process of computing the shape derivatives of cost functions. In Sect. 9.5, a state determination problem will be defined using a Poisson problem using the variation rules for functions shown in Sect. 9.4. The solution to this problem is used in Sect. 9.6 to define a general cost function which is then used to define a shape optimization problem. The existence of a solution to the shape optimization problem of this is shown in Sect. 9.7. In Sect. 9.8, the methods for obtaining the Fréchet derivatives of cost functions shown in Sect. 7.5 are followed in order to show the methods to obtain shape derivatives and second-order derivatives of cost functions with respect to a domain variation. In this case, we focus on the fact that we can think of two methods: one using formulae based on the shape derivative of a function, and another using formulae based on the partial shape derivative of a function. As a result, it becomes clear that whichever method is used, the shape gradients of the cost functions do not have enough regularity to be able to define the following domain.

Even if the shape gradients have insufficient regularities, by applying the abstract gradient method or the abstract Newton method shown in Sect. 7.6 to the shape optimization problems, a gradient method and Newton method with the functionality to regularize the shape derivatives of cost functions can be defined.

In Sect. 9.9, their abstract definitions and several methods for specifying these are introduced. In Sect. 9.10, algorithms will be considered. However, the basic structures are as per the algorithms shown in Sect. 3.7. The error evaluation of the numerical solutions obtained using these algorithms is shown in Sect. 9.11. Here, the results from the error estimations of numerical analyses shown in Sect. 6.6 will be used.

Once we look at the range of solutions with respect to the shape optimization problem of a Poisson problem, the shape derivatives of cost functions with respect to a mean compliance minimization problem of a linear elastic body will be sought in Sect. 9.12. Furthermore, in Sect. 9.13, the mean flow resistance minimization problem of a Stokes flow field will be used as an example to obtain the shape derivatives of the cost functions. The conditions of optimality using these shape derivatives can be seen matching the conditions of optimality with respect to the mean compliance minimization problem for a one-dimensional linear elastic body shown in Sect. 1.1 and the mean flow resistance minimization problem for a one-dimensional branched Stokes flow field shown in Sect. 1.3. Moreover, in Sects. 9.12.5 and 9.13.5, numerical examples with respect to these simple problems will be shown.

9.1 Set of Domain Variations and Definition of Shape Derivatives

In order to construct a shape optimization problem of domain variation type, let us define the admissible set of design variables. Moreover, the Fréchet derivatives of functions and functionals defined in a moving domain with respect to domain variation will be referred to as shape derivatives. These definitions will be shown in this section.

9.1.1 Initial Domain

Referring to Fig. 9.2, $\Omega_0 \subset \mathbb{R}^d$ is taken to be a $d \in \{2, 3\}$ -dimensional Lipschitz domain (Sect. A.5) representing the initial domain. In this chapter, we will assume that this boundary $\partial\Omega_0$ is also $H^2 \cap C^{0,1}$ class. Here, a boundary of $H^{k+2} \cap C^{k,1}$ class ($k \in \{0, 1, 2, \dots\}$) is defined as that the function ϕ defined in Definition A.5.2 (C^k class domain) belongs to $H^{k+2}(B(x, \alpha); \mathbb{R}^d) \cap C^{k,1}(B(x, \alpha); \mathbb{R}^d)$ ($H^{k+3/2}(\partial\Omega_0 \cap B(x, \alpha); \mathbb{R}^d) \cap C^{k,1}(\partial\Omega_0 \cap B(x, \alpha); \mathbb{R}^d)$ on boundary). Moreover, hereafter, we denote $H^{k+2}(\Omega_0; \mathbb{R}^d) \cap C^{k,1}(\Omega_0; \mathbb{R}^d)$ as $H^{k+2} \cap C^{k,1}(\Omega_0; \mathbb{R}^d)$.

It is assumed that Ω_0 is given. With respect to the boundary $\partial\Omega_0$ of the initial domain, $\Gamma_{D0} \subset \partial\Omega_0$ is taken to be a Dirichlet boundary and $\Gamma_{N0} =$

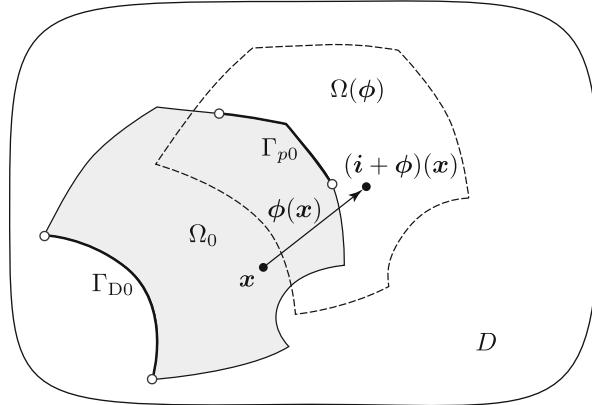


Fig. 9.2 Domain variation (displacement) $\phi : D \rightarrow \mathbb{R}^d$

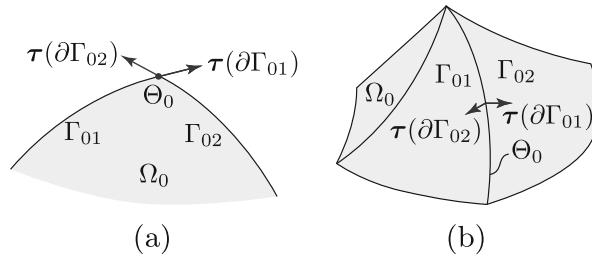


Fig. 9.3 Set of corner points (when $d = 2$) or edges (when $d = 3$) $\Theta_0 = \partial\Gamma_{01} \cap \partial\Gamma_{02}$ on boundary $\Gamma_0 = \Gamma_{01} \cup \Gamma_{02} \cup (\partial\Gamma_{01} \cap \partial\Gamma_{02}) \subset \partial\Omega_0$ and outward facing tangent $\tau(\partial\Gamma_{01})$ and $\tau(\partial\Gamma_{02})$ on $\partial\Gamma_{01}$ and $\partial\Gamma_{02}$, respectively. (a) $\Omega_0 \subset \mathbb{R}^2$. (b) $\Omega_0 \subset \mathbb{R}^3$

$\partial\Omega_0 \setminus \bar{\Gamma}_{D0}$ a Neumann boundary. Moreover, the notation for the set $(\bar{\cdot})$ is to represent a closure. Moreover, in this chapter, homogeneous Neumann boundaries and inhomogeneous Neumann boundaries will be distinguished from one another, and the inhomogeneous Neumann boundary of the initial domain will be written as $\Gamma_{p0} \subset \Gamma_{N0}$. Furthermore, we assume that the integrands used in the boundary integrals in the $m + 1$ cost functions f_0 (object cost function) and f_1, \dots, f_m (constraint cost functions) to be defined by Eq. (9.6.1) later will be denoted as η_{Ni} with respect to $i \in \{0, 1, \dots, m\}$ and these will be non-zero on $\Gamma_{\eta i 0} \subset \Gamma_{N0}$. If Γ_{p0} and $\Gamma_{\eta i 0}$ are assumed to vary, these boundaries are piecewise $H^3 \cap C^{1,1}$, and when $d = 3$, boundaries $\partial\Gamma_{p0}$ or $\partial\Gamma_{\eta i 0}$ are assumed to be $H^2 \cap C^{0,1}$ class. These hypotheses will be needed to guarantee appropriate regularity of shape derivatives of cost functions obtained on these boundaries. Moreover, when their boundaries are denoted as Γ_0 (Γ_0 is an open set excluding $\partial\Gamma_0$) as shown in Fig. 9.3, the set of corner points (when $d = 2$) or edges (when $d = 3$) on Γ_0 is denoted as Θ_0 , and

edges included in Θ_0 (when $d = 3$) are assumed to be $H^2 \cap C^{0,1}$ class. For $\Gamma(\cdot)$, the notation $\Theta(\cdot)$ is used.

9.1.2 Sets of Domain Variations

Let us define a domain after Ω_0 is perturbed. \mathbf{i} will represent the identity mapping. In this case, the domain after Ω_0 is perturbed is assumed to be formed by a continuous bijective mapping $\mathbf{i} + \boldsymbol{\phi} : \Omega_0 \rightarrow \mathbb{R}^d$ as $(\mathbf{i} + \boldsymbol{\phi})(\Omega_0) = \{(\mathbf{i} + \boldsymbol{\phi})(\mathbf{x}) \mid \mathbf{x} \in \Omega_0\}$. In other words, $\boldsymbol{\phi}$ is to represent the displacement in the domain mapping. Since the domain $(\mathbf{i} + \boldsymbol{\phi})(\Omega_0)$ is formed by $\boldsymbol{\phi}$, it is denoted by $\Omega(\boldsymbol{\phi})$. Similarly, with respect to an initial domain or boundary $(\cdot)_0$, $(\cdot)(\boldsymbol{\phi})$ represents $\{(\mathbf{i} + \boldsymbol{\phi})(\mathbf{x}) \mid \mathbf{x} \in (\cdot)_0\}$.

When the design variable $\boldsymbol{\phi}$ is selected as above, even though the domain of $\boldsymbol{\phi}$ is fixed at Ω_0 , the domain of the solution to a state determination problem varies with the domain variation. Such a situation is not expected to occur in a general function optimization problem. However, from the Calderón extension theorem (Theorem 4.4.4), if the domain of $\boldsymbol{\phi}$ is expanded to $D \subset \mathbb{R}^d$ large enough via Theorem 4.4.4, the conditions for ordinary function optimization problems are satisfied.

Hence, under conditions satisfying the assumption (with respect to $p > 1$, $\boldsymbol{\phi} \in W^{1,p}(\Omega_0; \mathbb{R}^d)$) of Theorem 4.4.4, we will expand the domain of $\boldsymbol{\phi}$ from Ω_0 to a bounded domain $D \subset \mathbb{R}^d$ large enough. Furthermore, since we will be considering the gradient method on a function space later, the function space containing the design variable $\boldsymbol{\phi}$ needs to be a Hilbert space. Hence in this chapter, the linear space of design variables is defined as

$$X = \left\{ \boldsymbol{\phi} \in H^1(D; \mathbb{R}^d) \mid \boldsymbol{\phi} = \mathbf{0}_{\mathbb{R}^d} \text{ on } \partial D \cup \bar{\Omega}_{C0} \right\}, \quad (9.1.1)$$

where $\bar{\Omega}_{C0} \subset \bar{\Omega}_0$ represents a boundary or closure of the domain in which the domain variation is constrained by a design demand. In continuing discussions in this chapter, it will be viewed as $\bar{\Omega}_{C0} = \emptyset$ (in other words, $X = H^1(D; \mathbb{R}^d)$). If it is needed that the measure of $\bar{\Omega}_{C0}$ is assumed to have a certain positive value, its condition will be clearly presented.

However, when $\boldsymbol{\phi}$ is taken to be an element of X , there is no guarantee that $\Omega(\boldsymbol{\phi})$ is a Lipschitz domain. In order to become a Lipschitz domain, $\boldsymbol{\phi}$ has to be an element of $C^{0,1}(D; \mathbb{R}^d)$. To repeat domain variations under the same conditions, the condition $(\Gamma_{p0} \cup \Gamma_{\eta00} \cup \Gamma_{\eta10} \cup \dots \cup \Gamma_{\eta m0} \setminus \bar{\Omega}_{C0})$ belongs to a class of piecewise $H^3 \cap C^{1,1}$ with respect to $\partial \Omega_0$ needs to be satisfied even for the perturbed boundary $\partial \Omega(\boldsymbol{\phi})$. Furthermore, to guarantee the existence of an optimum shape as shown in Sect. 9.7, the admissible set for $\boldsymbol{\phi}$ should be compact in X . Considering those conditions, one needs to take a linear space of $\boldsymbol{\phi}$, in which Fréchet derivatives of

functions and functionals with respect to domain variation can be defined, as

$$Y = \begin{cases} X \cap H^2 \cap C^{0,1}(D; \mathbb{R}^d) & (\tilde{\Gamma}_0 = \emptyset \text{ or } \tilde{\Gamma}_0 \subset \bar{\Omega}_{C0}), \\ X \cap H^3 \cap C^{1,1}(D; \mathbb{R}^d) & (\tilde{\Gamma}_0 \not\subset \bar{\Omega}_{C0}) \end{cases}, \quad (9.1.2)$$

where $\tilde{\Gamma}_0$ denotes $\Gamma_{p0} \cup \Gamma_{\eta00} \cup \Gamma_{\eta10} \cup \dots \cup \Gamma_{\eta m0}$ after Sect. 9.5 and a piecewise $H^3 \cap C^{1,1}$ class boundary before that section, and the admissible set of design variables as

$$\mathcal{D} = \left\{ \boldsymbol{\phi} \in Y \left| \begin{array}{l} |\boldsymbol{\phi}|_{C^{0,1}(D; \mathbb{R}^d)} \leq \sigma, \\ \|\boldsymbol{\phi}\|_{H^2 \cap C^{0,1}(D; \mathbb{R}^d)} \leq \beta \quad (\tilde{\Gamma}_0 = \emptyset \text{ or } \tilde{\Gamma}_0 \subset \bar{\Omega}_{C0}), \\ \|\boldsymbol{\phi}\|_{H^3 \cap C^{1,1}(D; \mathbb{R}^d)} \leq \beta \quad (\tilde{\Gamma}_0 \not\subset \bar{\Omega}_{C0}) \end{array} \right. \right\}. \quad (9.1.3)$$

Here, let $\|\boldsymbol{\phi}\|_{H^2 \cap C^{0,1}(D; \mathbb{R}^d)}$ be defined as $\max \left\{ \|\boldsymbol{\phi}\|_{H^2(D; \mathbb{R}^d)}, \|\boldsymbol{\phi}\|_{C^{0,1}(D; \mathbb{R}^d)} \right\}$, and $\sigma \in (0, 1)$ and β be positive constants. Norm $|\cdot|_{C^{0,1}(D; \mathbb{R}^d)} = \|\cdot\|_{C^{0,1}(D; \mathbb{R}^d)} - \|\cdot\|_{C(D; \mathbb{R}^d)}$ represents the Lipschitz constant (see (4.3.2)). The condition $|\boldsymbol{\phi}|_{C^{0,1}(D; \mathbb{R}^d)} \leq \sigma$ represents that $\mathbf{i} + \boldsymbol{\phi}$ and its inverse mapping $(\mathbf{i} + \boldsymbol{\phi})^{-1}$ become Lipschitz mappings (bi-Lipschitz mappings) on Ω_0 and $\Omega(\boldsymbol{\phi})$ respectively (cf. [97, Proposition 1.41, p. 23], [115]). Indeed, if this condition is satisfied, $\mathbf{i} + \boldsymbol{\phi}$ is a surjective Lipschitz mapping on Ω_0 . Moreover, using

$$\begin{aligned} \|(\mathbf{i} + \boldsymbol{\phi})(\mathbf{x}_0) - (\mathbf{i} + \boldsymbol{\phi})(\mathbf{y}_0)\|_{\mathbb{R}^d} &\geq \|\mathbf{x}_0 - \mathbf{y}_0\|_{\mathbb{R}^d} - \|\boldsymbol{\phi}(\mathbf{x}_0) - \boldsymbol{\phi}(\mathbf{y}_0)\|_{\mathbb{R}^d} \\ &\geq (1 - \sigma) \|\mathbf{x}_0 - \mathbf{y}_0\|_{\mathbb{R}^d} \end{aligned}$$

for arbitrary $\mathbf{x}_0, \mathbf{y}_0 \in \Omega_0$, and that when $(\mathbf{i} + \boldsymbol{\phi})(\mathbf{x}_0) = (\mathbf{i} + \boldsymbol{\phi})(\mathbf{y}_0)$, we obtain that $\mathbf{x}_0 = \mathbf{y}_0$, $\mathbf{i} + \boldsymbol{\phi}$ becomes injective. Then, there exists $(\mathbf{i} + \boldsymbol{\phi})^{-1}$ with which

$$\|\mathbf{x}_1 - \mathbf{y}_1\|_{\mathbb{R}^d} \geq (1 - \sigma) \|(\mathbf{i} + \boldsymbol{\phi})^{-1}(\mathbf{x}_1) - (\mathbf{i} + \boldsymbol{\phi})^{-1}(\mathbf{y}_1)\|_{\mathbb{R}^d}$$

holds for arbitrary $\mathbf{x}_1, \mathbf{y}_1 \in \Omega(\boldsymbol{\phi})$. This inequality shows that $(\mathbf{i} + \boldsymbol{\phi})^{-1}$ is a Lipschitz mapping on $\Omega(\boldsymbol{\phi})$. On the other hand, that \mathcal{D} is a compact set in X is assured by the Rellich–Kondrachov compact embedding theorem (Theorem 4.4.15).

In future discussions, we assume that $\boldsymbol{\phi}$ is in the interior of \mathcal{D} ($\boldsymbol{\phi} \in \mathcal{D}^\circ$), and a domain perturbed via domain variation $\boldsymbol{\varphi} \in Y$ such as in Fig. 9.4 will be denoted by $\Omega(\boldsymbol{\phi} + \boldsymbol{\varphi})$. In the condition satisfying $\boldsymbol{\varphi} \in \mathcal{D}$, the perturbed domain will be defined as

$$(\Omega(\boldsymbol{\phi}))(\boldsymbol{\varphi}) = ((\mathbf{i} + \boldsymbol{\varphi}) \circ (\mathbf{i} + \boldsymbol{\phi}))(\Omega_0),$$

where \circ denotes the composite mapping. However, in future discussions, we will define the shape derivatives of function and functional as bounded linear operators

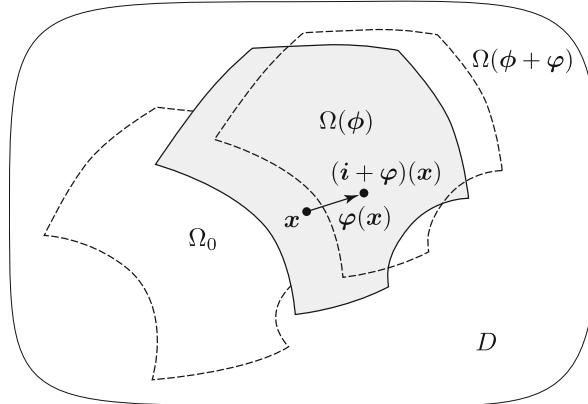


Fig. 9.4 Domain variation $\varphi \in Y$ from $\Omega(\phi)$

of $\varphi \in X$ (Definitions 9.1.1, 9.1.3, 9.1.4). In the Fréchet derivatives of functions and functionals with respect to $\varphi \in X$, $(\Omega(\phi))(\varphi)$ is linearized as $\Omega(\phi + \varphi)$. Then, in this chapter, we assume that φ is originally an element of X , check that φ obtained by the proposed methods belongs to Y based on the problem setting and solution used, and confirm that $\phi + \epsilon\varphi$ (ϵ is a positive constant) belongs to \mathcal{D} .

9.1.3 Definitions of Shape Derivatives

For problems involving varying domains, the functions and integrals also vary. Here let us define their shape derivatives.

Let $\phi_0 \in \mathcal{D}^\circ$ be given. For ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, let $\varphi \in Y$ be an arbitrary variation of $\Omega(\phi)$. When the domain varies from $\Omega(\phi)$ to $\Omega(\phi + \varphi)$, the function defined on it is also assumed to change. In this case, we write the function at ϕ as $u(\phi)$ and the value at a point x on the expanded domain D of $\Omega(\phi)$ as $u(\phi)(x)$. We use this notation to define the shape derivative of a function in the following way.

Definition 9.1.1 (Shape Derivative of a Function) For all ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, consider a function $u : B \rightarrow L^2(D; \mathbb{R})$. The value of $u(\phi)$ at $x \in D$ will be written as $u(\phi)(x)$. If there exists a bounded linear operator $u'(\phi) : Y \rightarrow L^2(D; \mathbb{R})$ which satisfies

$$\lim_{\|\varphi\|_Y \rightarrow 0} \frac{\|u(\phi + \varphi)(x + \varphi(x)) - u(\phi)(x) + u'(\phi)[\varphi](x)\|_{L^2(D; \mathbb{R})}}{\|\varphi\|_X} = 0$$

with respect to an arbitrary $\varphi \in Y$, and $u'(\phi)[\cdot] : X \rightarrow L^2(D; \mathbb{R})$ is also a bounded linear operator, $u'(\phi)[\varphi]$ is referred to as the shape derivative at $\phi \in B$ of u . When $u'(\phi)[\varphi]$ exists for all $\phi \in B$ and belongs to $C(B; \mathcal{L}(X; L^2(D; \mathbb{R})))$, we write $u \in C_{S'}^1(B; L^2(D; \mathbb{R}))$. \square

In Definition 9.1.1, we remark the following.

Remark 9.1.2 (Shape Derivative) In Definition 9.1.1, at the same time that $u'(\phi)[\cdot]$ is a bounded linear operator on Y , it is assumed that it is also a bounded linear operator in X . When Y is compactly embedded in X , a bounded linear operator in X is automatically a bounded linear operator in Y (Practice 4.4). The reason to define it in such a way is that, in general, $\Omega(\phi + \varphi)$ varied by an arbitrary $\varphi \in X$ is not well-defined. Also, by defining it such a way, stronger regularity is required than when the shape derivatives are defined as bounded linear operators in Y . In some types of definitions regarding shape derivatives shown later, the condition to be a bounded linear operator in X will be needed for the same reason. \square

In continuum mechanics, $u(\phi + \varphi)(x + \varphi(x))$ in Definition 9.1.1 is called the Lagrangian description of $u(\phi)(x)$ and $u'(\phi)[\varphi]$ is called the material derivative.

Figure 9.5a shows $u'(\phi)[\varphi]$. Here, even if $u \in L^2(D; \mathbb{R})$ is a discontinuous function, if φ is a continuous function, it is apparent that $u'(\phi)[\varphi]$ can be defined.

Next, let us think about the derivative of $u(\phi + \varphi)(x)$ when a point x is fixed on the expanded domain D of $\Omega(\phi)$ in the case of the perturbed domain. The Fréchet derivative of u with respect to an arbitrary variation φ is called the partial shape derivative of a function and is defined as follows.

Definition 9.1.3 (Partial Shape Derivative of a Function) For all ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, consider a function $u : B \rightarrow C^{0,1}(D; \mathbb{R})$. The value of $u(\phi)$ at $x \in D$ will be written as $u(\phi)(x)$. With respect to an arbitrary $\varphi \in Y$, when a bounded linear operator $u^*(\phi)[\cdot] : Y \rightarrow C^{0,1}(D; \mathbb{R})$ which satisfies that

$$\lim_{\|\varphi\|_Y \rightarrow 0} \frac{\|u(\phi + \varphi)(x) - u(\phi)(x) + u^*(\phi)[\varphi](x)\|_{C^{0,1}(D; \mathbb{R})}}{\|\varphi\|_X} = 0$$

for almost every $x \in D$ exists and when $u^*(\phi)[\cdot] : X \rightarrow C^{0,1}(D; \mathbb{R})$ is also a bounded linear operator, $u^*(\phi)[\varphi]$ is called the partial shape derivative of u at $\phi \in B$. When $u^*(\phi)[\varphi]$ exists for all $\phi \in B$ and belongs to $C(B; \mathcal{L}(X; C^{0,1}(D; \mathbb{R})))$, we write $u \in C_{S^*}^1(B; C^{0,1}(D; \mathbb{R}))$. \square

In Definition 9.1.3, $u(\phi + \varphi)(x)$ is called the Euler description of $u(\phi)(x)$ in continuum mechanics, and $u^*(\phi)[\varphi]$ is called the spatial derivative.

Figure 9.5b shows $u^*(\phi)[\varphi]$. Here, it should be noted that since $u \in H^1(D; \mathbb{R})$ is a continuous function, the definition of $u^*(\phi)[\varphi]$ is valid. In reality, looking at Fig. 9.5a, if u is a discontinuous function, $u^*(\phi)[\varphi]$ is not defined for x such that a discontinuity of u crosses in the domain variation due to φ .

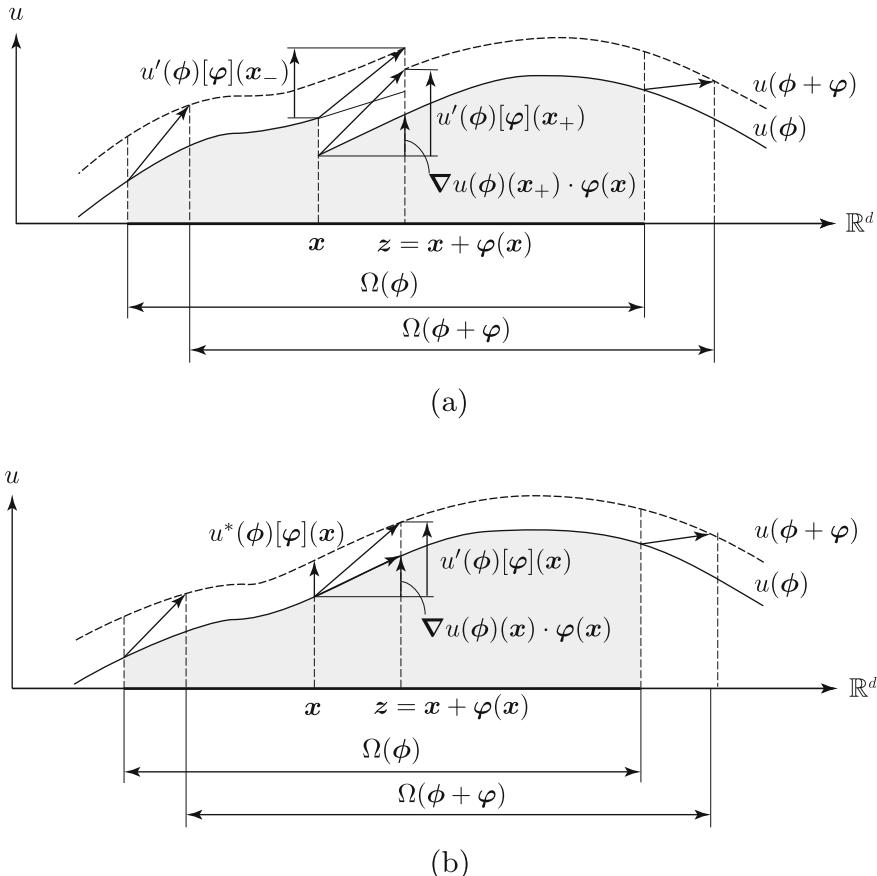


Fig. 9.5 The function $u(\phi)$ varying with domain. **(a)** When $u(\phi)$ is a discontinuous function. **(b)** When $u(\phi)$ is a continuous function

Moreover, when $u \in C_{S^*}^1(B; C^{0,1}(D; \mathbb{R}))$,

$$\begin{aligned}
 u(\phi + \varphi)(x + \varphi(x)) &= u(\phi + \varphi)(x) + \nabla u(\phi) \cdot \varphi + o(\|\varphi(x)\|_X) \\
 &= u(\phi)(x) + u^*(\phi)[\varphi](x) + \nabla u(\phi) \cdot \varphi + o(\|\varphi(x)\|_X)
 \end{aligned}$$

holds. Then, we have

$$u'(\phi)[\varphi] = u^*(\phi)[\varphi] + \nabla u(\phi) \cdot \varphi. \quad (9.1.4)$$

with respect to arbitrary $\varphi \in X$. Here $(\partial(\cdot)/x_1, \dots, \partial(\cdot)/x_d)^\top$ is denoted as $\nabla(\cdot)$ with respect to $\mathbf{x} = (x_i)_{i \in \{1, \dots, d\}} \in \mathbb{R}^d$. The left-hand side of Eq. (9.1.4) becomes $u'(\phi)[\cdot] : X \rightarrow L^2(D; \mathbb{R})$. Then, $u \in C_{S'}^1(B; L^2(D; \mathbb{R}))$ holds.

Furthermore, the shape derivative of a functional defined on a perturbed domain will be defined as follows. In this chapter, we use the notation $\nabla_z = (\partial(\cdot)/z_1, \dots, \partial(\cdot)/z_d)^\top$ with respect to $z = (\phi + \varphi)(\mathbf{x})$. Moreover, let $\mathbf{v}(\phi)$ be the outward unit normal defined on the boundary $\partial\Omega(\phi)$, $\partial_v(\cdot) = \mathbf{v}(\phi) \cdot \nabla(\cdot)$, $\mu = (\phi + \varphi)(\mathbf{v})$ be the outward unit normal on $\partial\Omega(\phi + \varphi)$, and $\partial_\mu(\cdot) = \mu \cdot \nabla_z(\cdot)$. Furthermore, the dual space of X is denoted by X' .

Definition 9.1.4 (Shape Derivative of a Functional) For all ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, let $u : B \rightarrow \mathcal{U} = H^3 \cap C^{1,1}(D; \mathbb{R})$ be given, and $h_0 \in C^1(\mathbb{R} \times \mathbb{R}^d; \mathbb{R})$ and $h_1 \in C^1(\mathbb{R} \times \mathbb{R}; \mathbb{R})$ be defined for $(u, \nabla u, \partial_v u) \in \mathcal{U} \times \mathcal{G} \times \mathcal{G}_{\Gamma(\phi)}$ ($\mathcal{G} = \{\nabla u \mid u \in \mathcal{U}\}$, $\mathcal{G}_{\Gamma(\phi)} = \{\partial_v u|_{\Gamma(\phi)} \mid u \in \mathcal{U}\}$) ($\Gamma(\phi) \subseteq \partial\Omega(\phi)$ is piecewise $H^3 \cap C^{1,1}$) as

$$h_0(u, \nabla u), h_{0u}(u, \nabla u) \in L^2(D; \mathbb{R}), \quad h_{0\nabla u}(u, \nabla u) \in L^2(D; \mathbb{R}^d),$$

$$h_1(u, \partial_v u), h_{1u}(u, \partial_v u), h_{1\partial_v u}(u, \partial_v u) \in H^1(D; \mathbb{R}).$$

With respect to an arbitrary $\varphi \in Y$, let

$$\begin{aligned} & f(\phi + \varphi, u(\phi + \varphi), \nabla_z u(\phi + \varphi), \partial_\mu u(\phi + \varphi)) \\ &= \int_{\Omega(\phi + \varphi)} h_0(u(\phi + \varphi)(z), \nabla_z u(\phi + \varphi)(z)) dz \\ &+ \int_{\Gamma(\phi + \varphi)} h_1(u(\phi + \varphi)(z), \partial_\mu u(\phi + \varphi)(z)) d\zeta. \end{aligned} \quad (9.1.5)$$

Here, $\Gamma(\phi)$ is taken to be the partial set of $\partial\Omega(\phi)$ (allowing $\Gamma(\phi) = \partial\Omega(\phi)$). Moreover, dz and $d\zeta$ represent infinitesimal measures used in domain and boundary integrals over $\Omega(\phi + \varphi)$. In this case, if a bounded linear functional $f'(\phi, u(\phi), \nabla u(\phi), \partial_v u(\phi))[\cdot] : Y \rightarrow \mathbb{R}$ satisfies

$$\begin{aligned} & f(\phi + \varphi, u(\phi + \varphi), \nabla_z u(\phi + \varphi), \partial_\mu u(\phi + \varphi)) \\ &= f(\phi, u(\phi), \nabla u(\phi), \partial_v(\phi)) \\ &+ f'(\phi, u(\phi), \nabla u(\phi), \partial_v(\phi))[\varphi] + o(\|\varphi\|_X) \end{aligned}$$

and $f'(\phi, u(\phi), \nabla u(\phi), \partial_v u(\phi))[\cdot] : X \rightarrow \mathbb{R}$ is also a bounded linear functional, in other words, there exists a $\mathbf{g}(\phi) \in X'$ such that $f'(\phi, u(\phi), \nabla u(\phi), \partial_v u(\phi))[\varphi] = \langle \mathbf{g}(\phi), \varphi \rangle$, f is said to be shape differentiable at ϕ , and $\mathbf{g}(\phi)$ is called the shape gradient of f . Moreover, when there exists $f'(\phi, u(\phi), \nabla u(\phi), \partial_v(\phi))[\varphi]$ for all $\phi \in B$ and those are in $C(B; \mathcal{L}(X; \mathbb{R}))$, it is expressed as $f \in C_S^1(B; \mathbb{R})$.

Furthermore, if with respect to an arbitrary $\varphi_1, \varphi_2 \in Y$, a bounded bilinear functional $f''(\phi, u(\phi), \nabla u(\phi), \partial_v(\phi))[\varphi_1, \varphi_2] : Y \times Y \rightarrow \mathbb{R}$ satisfies

$$\begin{aligned} & \left\langle g(\phi + \varphi_2), \varphi_1 \circ (i + \varphi_2)^{-1} \right\rangle \\ &= \langle g(\phi), \varphi_1 \rangle + f''(\phi, u(\phi), \nabla u(\phi), \partial_v(\phi))[\varphi_1, \varphi_2] \\ &+ o(\|\varphi_1\|_X, \|\varphi_2\|_X) \end{aligned}$$

and $f''(\phi, u(\phi), \nabla u(\phi), \partial_v(\phi))[\varphi_1, \varphi_2] = h(\phi)[\varphi_1, \varphi_2] : X \times X \rightarrow \mathbb{R}$ is also a bounded bilinear functional, f is said to be second-order shape differentiable, and $h(\phi)[\varphi_1, \varphi_2]$ is called the second-order shape derivative or shape Hessian of f . In addition, with respect to all $\phi \in B$, if there exists a second-order shape derivative and $f''(\phi, u(\phi), \nabla u(\phi), \partial_v(\phi))[\varphi_1, \varphi_2] \in C(B; \mathcal{L}(X; \mathcal{L}(X; \mathbb{R})))$, then we write that $f \in C_S^2(B; \mathbb{R})$. \square

According to the definition of the second-order shape derivative, $f''(\phi, u(\phi), \nabla u(\phi), \partial_v(\phi))[\varphi_1, \varphi_2]$ can be divided into two parts as [153]

$$\begin{aligned} & f''(\phi, u(\phi), \nabla u(\phi), \partial_v(\phi))[\varphi_1, \varphi_2] \\ &= (f')'(\phi, u(\phi), \nabla u(\phi), \partial_v(\phi))[\varphi_1, \varphi_2] + \langle g(\phi), t(\varphi_1, \varphi_2) \rangle. \end{aligned} \quad (9.1.6)$$

Here, the summands on the right side of the above equation are respectively given as

$$\begin{aligned} & (f')'(\phi, u(\phi), \nabla u(\phi), \partial_v(\phi))[\varphi_1, \varphi_2] \\ &= \lim_{\|\varphi_2\|_X \rightarrow 0} \frac{1}{\|\varphi_2\|_X} (\langle g(\phi + \varphi_2), \varphi_1 \rangle - \langle g(\phi), \varphi_1 \rangle) \end{aligned} \quad (9.1.7)$$

$$\begin{aligned} & \langle g(\phi), t(\varphi_1, \varphi_2) \rangle \\ &= \lim_{\|\varphi_2\|_X \rightarrow 0} \frac{1}{\|\varphi_2\|_X} \left\langle g(\phi + \varphi_2), \varphi_1 \circ (i + \varphi_2)^{-1} - \varphi_1 \right\rangle. \end{aligned} \quad (9.1.8)$$

Equation (9.1.7) represents the derivative of $g(\phi + \varphi_2)$ by the variation of φ_2 , and commonly appears in calculations of the second-order derivative of a functional in optimization problems. On the other hand, Eq. (9.1.8) is a specific term in shape optimization problems to correct the variation of φ_1 by φ_2 . The term $\varphi_1 \circ (i + \varphi_2)^{-1} - \varphi_1$ in Eq. (9.1.8) represents the variation of φ_1 by the inverse

mapping of $\mathbf{i} + \boldsymbol{\varphi}_2$. When only this item is calculated, we have

$$\begin{aligned} t(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) &= \lim_{\|\boldsymbol{\varphi}_2\|_X \rightarrow 0} \frac{1}{\|\boldsymbol{\varphi}_2\|_X} \left(\boldsymbol{\varphi}_1 \circ (\mathbf{i} + \boldsymbol{\varphi}_2)^{-1} - \boldsymbol{\varphi}_1 \right) \\ &= -\left(\boldsymbol{\varphi}_2 \cdot \nabla \boldsymbol{\varphi}_1^\top \right)^\top = -\left(\nabla \boldsymbol{\varphi}_1^\top \right)^\top \boldsymbol{\varphi}_2. \end{aligned} \quad (9.1.9)$$

The above relation can be obtained in the following way. We notice that the transfer vector of the coordinate system linearizing the inverse mapping of $\mathbf{i} + \boldsymbol{\varphi}_2$ becomes $-\boldsymbol{\varphi}_2$, and the varying function is $\boldsymbol{\varphi}_1$. In Eq. (9.1.4), replacing u by $\boldsymbol{\varphi}_1^\top$ and putting $\boldsymbol{\varphi}_1^*(\boldsymbol{\phi})[-\boldsymbol{\varphi}_2] = \mathbf{0}_{\mathbb{R}^d}$, $\boldsymbol{\varphi}_1'(\boldsymbol{\phi})(\boldsymbol{\phi})[-\boldsymbol{\varphi}_2]$ gives the right-hand side of Eq. (9.1.9).

It should be noted that Eq. (9.1.9) holds if $\nabla \boldsymbol{\varphi}^\top$ or $\nabla \cdot \boldsymbol{\varphi}$ is not used in the shape derivative $\langle \mathbf{g}(\boldsymbol{\phi}), \boldsymbol{\varphi} \rangle$. In the cases that $\nabla \boldsymbol{\varphi}^\top$ or $\nabla \cdot \boldsymbol{\varphi}$ is used, those calculations for $\nabla \boldsymbol{\varphi}^\top$ and $\nabla \cdot \boldsymbol{\varphi}$ will be given in Eq. (9.3.11). In that situation, the inverse mapping of $\mathbf{i} + \boldsymbol{\varphi}_2$ is applied to ∇ too.

9.2 Shape Derivatives of Jacobi Determinants

Since the domain variation and shape derivatives of functions and functionals have been defined, let us use them to find the shape derivative of the Jacobi determinant (Jacobian) and the inverse matrix of Jacobi matrix (Jacobi inverse matrix) with respect to domain variation $\boldsymbol{\varphi} \in Y$. These are used when seeking the formulae for the shape derivatives of functionals.

Fix $\boldsymbol{\phi}_0 \in \mathcal{D}^\circ$. For $\boldsymbol{\phi}$ in a neighborhood $B \subset Y$ of $\boldsymbol{\phi}_0 \in \mathcal{D}^\circ$, consider an arbitrary domain variation $\boldsymbol{\varphi} \in Y$ from $\Omega(\boldsymbol{\phi})$. In this case, the Jacobi matrix and Jacobi determinant (Jacobian) with respect to the mapping $\mathbf{i} + \boldsymbol{\varphi}$ are expressed as

$$\mathbf{F}(\boldsymbol{\varphi}) = \mathbf{I} + \left(\nabla \boldsymbol{\varphi}^\top \right)^\top, \quad (9.2.1)$$

$$\omega(\boldsymbol{\varphi}) = \det \mathbf{F}(\boldsymbol{\varphi}), \quad (9.2.2)$$

where \mathbf{I} represents the unit matrix.³ In this case, $\omega(\boldsymbol{\varphi})$ becomes a function which gives $dz = \omega(\boldsymbol{\varphi}) dx$ with respect to an infinitesimal measure dz on $\Omega(\boldsymbol{\phi} + \boldsymbol{\varphi})$ corresponding to the infinitesimal measure dx on $\Omega(\boldsymbol{\phi})$. Here, taking up two types of Jacobi determinants defined on the domain and the boundary, let us look at their shape derivatives.

³Although it is usually written as $\mathbf{F}(\mathbf{i} + \boldsymbol{\varphi})$, following the notation for a deformation gradient tensor in elasticity theory, $\mathbf{F}(\boldsymbol{\varphi})$ is used.

9.2.1 Shape Derivatives of Domain Jacobi Determinant and Domain Jacobi Inverse Matrix

Firstly, the shape derivative of $\omega(\phi)$ defined on Eq. (9.2.2) at $\phi_0 = \mathbf{0}_{\mathbb{R}^d}$ is given in the following way.

Proposition 9.2.1 (Derivative of Domain Jacobi Determinant) *For ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, we have*

$$\omega'(\phi_0)[\phi] = \nabla \cdot \phi$$

with respect to an arbitrary $\phi \in Y$. Moreover, $\omega'(\phi_0)[\phi]$ also belongs to $C(B; \mathcal{L}(X; L^2(D; \mathbb{R})))$. \square

Proof For $x \in D$, we have

$$\begin{aligned} \omega(\phi) &= \det \left(\mathbf{I} + (\nabla \phi^\top)^\top \right) = \det \begin{pmatrix} 1 + \varphi_{1,1} & \cdots & \varphi_{1,d} \\ \vdots & \ddots & \vdots \\ \varphi_{d,1} & \cdots & 1 + \varphi_{d,d} \end{pmatrix} \\ &= 1 + \nabla \cdot \phi + \sum_{(i,j) \in \{1, \dots, d\}^2} o\left(\|\varphi_{i,j}\|_{L^2(D; \mathbb{R})}\right). \end{aligned} \quad \square$$

Moreover, the shape derivative of the Jacobi inverse matrix $\mathbf{F}^{-\top}(\phi)$ at $\phi_0 = \mathbf{0}_{\mathbb{R}^d}$ is as follows.

Proposition 9.2.2 (Derivative of Domain Jacobi Inverse Matrix) *For ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, we have*

$$\mathbf{F}^{-\top'}(\phi_0)[\phi] = -\nabla \phi^\top$$

with respect to an arbitrary $\phi \in Y$. Moreover, $\mathbf{F}^{-\top'}(\phi_0)[\phi]$ also belongs to $C(B; \mathcal{L}(X; L^2(D; \mathbb{R}^{d \times d})))$. \square

Proof For $x \in D$, using the differentiability of the inverse map $(i + \phi)^{-1}$, we have

$$\mathbf{F}^{-\top}(\phi) \left(\mathbf{I} + \nabla \phi^\top \right) = \mathbf{I}.$$

Taking its shape derivative with respect to ϕ at ϕ , we get

$$\mathbf{F}^{-\top'}(\phi_0)[\phi] + \mathbf{F}^{-\top}(\phi_0) \left(\nabla \phi^\top \right) = \mathbf{0}_{\mathbb{R}^{d \times d}}.$$

Since $\mathbf{F}^{-\top}(\phi_0) = \mathbf{I}$, the proposition follows. \square

9.2.2 Shape Derivatives of Boundary Jacobi Determinant and the Normal

Next let us obtain the formulae for the shape derivatives relating to the Jacobi determinant on a boundary. In shape optimization problems of domain variation type, boundary integrals appear in the Lagrange functions of state determination problems and cost functions. Hence, when obtaining the shape derivatives of such boundary integrals, the shape derivatives of the boundary Jacobi determinant and the normal are needed.

Let us represent an infinitesimal measure on $\partial\Omega(\phi)$ by $d\gamma(\phi)$ and an outward unit normal by $v(\phi)$. Furthermore, the normal on a Lipschitz boundary is defined by the normal with respect to the graph defining the boundary as a graph in a local coordinate system around the boundary, and is assumed to be in $L^\infty(\partial\Omega(\phi); \mathbb{R}^d)$ [48, 114]. Here, we assume that $\partial\Omega(\phi)$ is piecewise $H^2 \cap C^{0,1}$ and $v(\phi) \in H^{1/2} \cap L^\infty(\partial\Omega(\phi); \mathbb{R}^d)$.

In this case, with respect to arbitrary $\varphi \in Y$, the relation

$$\varpi(\varphi) = \frac{d\gamma(\phi + \varphi)}{d\gamma(\phi)} = \omega(\varphi) v(\phi + \varphi) \cdot (F^{-\top}(\varphi) v(\phi)) \quad (9.2.3)$$

holds. Here, $\varpi(\varphi)$ denotes the Jacobi determinant for the boundary. This relationship can be obtained from the following proposition.

Proposition 9.2.3 (Nanson Formula) *For ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, let $\partial\Omega(\phi)$ be piecewise $H^2 \cap C^{0,1}$. For an arbitrary $\varphi \in Y$, the equation*

$$v(\phi + \varphi) d\gamma(\phi + \varphi) = \omega(\varphi) F^{-\top}(\varphi) v(\phi) d\gamma(\phi) \quad (9.2.4)$$

holds. Moreover, $\omega(\varphi) F^{-\top}(\varphi) v(\phi)$ belongs to $L^\infty(\partial\Omega(\phi); \mathbb{R})$. \square

Proof Let $dl(\phi) \in \mathbb{R}^d$ be an arbitrary vector satisfying $v(\phi) \cdot dl(\phi) > 0$ on $d\gamma(\phi)$ and $dl(\phi + \varphi)$ a vector obtained through the mapping $i + \varphi$. In this case, the relation

$$dl(\phi + \varphi) \cdot v(\phi + \varphi) d\gamma(\phi + \varphi) = \omega(\varphi) dl(\phi) \cdot v(\phi) d\gamma(\phi)$$

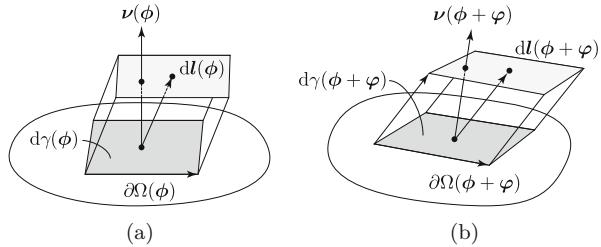
holds with respect to the volume of a parallelepiped shown in Fig. 9.6. Here, if $dl(\phi + \varphi) = F(\varphi) dl(\phi)$ is substituted into the equation above, one obtains

$$dl(\phi) \cdot (F^\top(\varphi) v(\phi + \varphi)) d\gamma(\phi + \varphi) = dl(\phi) \cdot (\omega(\varphi) v(\phi)) d\gamma(\phi).$$

Since $dl(\phi)$ is arbitrary, Eq. (9.2.4) follows. \square

Equation (9.2.4) can be obtained by using the Piola transformation giving the correspondence between second-order tensor functions defined over the deformed and initial domains [41, Theorem 1.7-1]. The Piola transformation $A(\varphi)$ of an

Fig. 9.6 Infinitesimal measures $d\gamma(\phi)$ and $d\gamma(\phi + \varphi)$. (a) Before variation. (b) After variation



arbitrary second-order tensor-valued function $A \in C^1(D; \mathbb{R}^{d \times d})$ with respect to an arbitrary $\varphi \in Y$ is defined as

$$A = \omega(\varphi) A(\varphi) F^{-\top}(\varphi),$$

where $\omega(\varphi) F^{-\top}(\varphi)$ is the cofactor matrix of the Jacobi matrix $F(\varphi)$. Letting $z = x + \varphi(x) = (i + \varphi)(x)$ on D be an admissible perturbation φ of a point x from $\Omega(\phi)$ after deformation, and ∇_z represents $\partial(\cdot)/\partial z$, we have

$$\begin{aligned} \int_{\Omega(\phi)} \nabla \cdot A dx &= \int_{\partial\Omega(\phi)} A \nu(\phi) d\gamma(\phi) \\ &= \int_{\Omega(\phi)} \nabla_z \cdot A(\varphi) \omega(\varphi) dx \\ &= \int_{\Omega(\phi+\varphi)} \nabla_z \cdot A(\varphi) dz \\ &= \int_{\partial\Omega(\phi+\varphi)} A(\varphi) \nu(\phi + \varphi) d\gamma(\phi + \varphi). \end{aligned}$$

Applying the Piola transformation to the above equality with respect to the boundary integral equation and putting $A(\varphi) = I$, we obtain Eq. (9.2.4).

Taking the inner product of both sides of Eq. (9.2.4) and $\nu(\phi + \varphi)$ leads to Eq. (9.2.3). Moreover, from the fact that $\nu(\phi + \varphi)$ is a unit vector in the direction of $F^{-\top}(\varphi) \nu(\phi)$, then by Eq. (9.2.4), the following holds:

$$\nu(\phi + \varphi) = \frac{F^{-\top}(\varphi) \nu(\phi)}{\|F^{-\top}(\varphi) \nu(\phi)\|_{\mathbb{R}^d}}. \quad (9.2.5)$$

Based on these relationships, the shape derivative of $\varpi(\varphi)$ of Eq. (9.2.3) can be obtained in Proposition 9.2.4. In the sequel, the tangent (Definition A.5.3) on $\partial\Omega(\phi)$ will be written as $\tau_1(\phi), \dots, \tau_{d-1}(\phi)$. On the Lipschitz boundary, the tangent is defined as a tangent on the graph defining the boundary as a graph of the local coordinate system near the boundary, in a similar way to the normal, and is assumed to be included in $L^\infty(\partial\Omega(\phi); \mathbb{R}^d)$. Moreover, $d - 1$ times of the mean curvature

(Definition A.5.5) (sum of principle curvatures) is given by $\kappa(\phi) = \nabla \cdot v(\phi)$ on a piecewise $C^{1,1}$ class boundary in a similar way to the derivative of the normal, and is assumed to be included in $L^\infty(\partial\Omega(\phi); \mathbb{R})$. Here, we assume that $\partial\Omega(\phi)$ is piecewise $H^3 \cap C^{1,1}$ and $\kappa(\phi) \in H^{1/2} \cap L^\infty(\partial\Omega(\phi); \mathbb{R})$. Moreover, $\nabla_\tau(\cdot) = (\tau_j(\phi) \cdot \nabla)_{j \in \{1, \dots, d-1\}}(\cdot) \in \mathbb{R}^{d-1}$ and $\varphi_\tau = (\tau_j(\phi) \cdot \varphi)_{j \in \{1, \dots, d-1\}} \in \mathbb{R}^{d-1}$. From now on, $v(\phi)$, $\tau_1(\phi), \dots, \tau_{d-1}(\phi)$ and $\kappa(\phi)$ are to be written simply as v , $\tau_1, \dots, \tau_{d-1}$ and κ .

Proposition 9.2.4 (Derivative of Boundary Jacobi Determinant) *For ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, let $\partial\Omega(\phi)$ be piecewise $H^2 \cap C^{0,1}$. In this case, we have the identity*

$$\varpi'(\phi_0)[\varphi] = (\nabla \cdot \varphi)_\tau = \nabla \cdot \varphi - v \cdot (\nabla \varphi^\top v) \quad (9.2.6)$$

with respect to an arbitrary $\varphi \in Y$, where $(\nabla \cdot \varphi)_\tau$ is defined by the right-hand side of Eq. (9.2.6). Furthermore, if $\partial\Omega(\phi)$ is a piecewise $H^3 \cap C^{1,1}$ boundary, one has

$$\varpi'(\phi_0)[\varphi] = \kappa v \cdot \varphi + \nabla_\tau \cdot \varphi_\tau. \quad (9.2.7)$$

Moreover, $\varpi'(\phi_0)[\varphi]$ belongs to $C(B; \mathcal{L}(Y; L^\infty(\partial\Omega(\phi); \mathbb{R})))$. \square

Proof From Eqs. (9.2.3) and (9.2.5), we have

$$\varpi(\varphi) = \omega(\varphi) \left\| F^{-\top}(\varphi) v \right\|_{\mathbb{R}^d}.$$

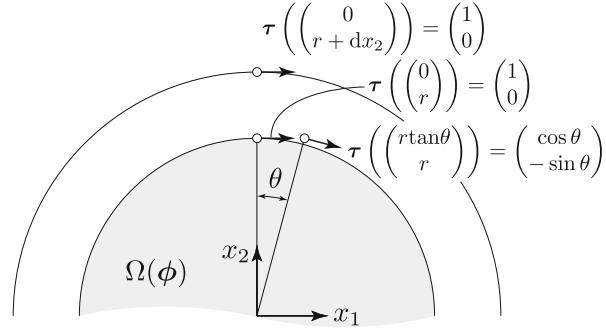
Equation (9.2.6) can be obtained from Propositions 9.2.1 and 9.2.2 as

$$\begin{aligned} \varpi'(\phi_0)[\varphi] &= \omega'(\phi_0)[\varphi] \left\| F^{-\top}(\phi_0) v \right\|_{\mathbb{R}^d} \\ &\quad + \omega(\phi_0) \left(F^{-\top}(\phi_0) v \right) \cdot \left(F^{-\top'}(\phi_0)[\varphi] v \right) \Big/ \left\| F^{-\top}(\phi_0) v \right\|_{\mathbb{R}^d} \\ &= \nabla \cdot \varphi - v \cdot (\nabla \varphi^\top v). \end{aligned}$$

Furthermore, if its boundary is piecewise $H^3 \cap C^{1,1}$, we can define $\kappa = \nabla \cdot v$ almost everywhere and write

$$\begin{aligned} \nabla \cdot \varphi &= \nabla \cdot \left\{ (v \cdot \varphi) v + \sum_{j \in \{1, \dots, d-1\}} (\tau_j \cdot \varphi) \tau_j \right\} \\ &= \partial_v(v \cdot \varphi) + \kappa(v \cdot \varphi) + \nabla_\tau \cdot \varphi_\tau. \end{aligned} \quad (9.2.8)$$

Fig. 9.7 Distribution of a tangent near a circle



Here, the notation $\nabla \cdot \tau_1 = 0, \dots, \nabla \cdot \tau_{d-1} = 0$ is used. This is because when $\Omega(\phi)$ is a circle (a two-dimensional domain) with radius r , such as in Fig. 9.7, one has

$$\nabla \cdot \tau_1 = \nabla \cdot \tau = \frac{\partial \tau_1}{\partial x_1} + \frac{\partial \tau_2}{\partial x_2} = \lim_{\theta \rightarrow 0} \frac{\cos \theta - 1}{r \tan \theta} = 0$$

at $x = (0, r)^T$. A similar relationship holds even when $\Omega(\phi)$ is a three-dimensional domain.

Moreover,

$$\begin{aligned} \mathbf{v} \cdot (\nabla \phi^\top \mathbf{v}) &= \mathbf{v} \cdot \left[\nabla \left\{ (\mathbf{v} \cdot \phi) \mathbf{v} + \sum_{j \in \{1, \dots, d-1\}} (\tau_j \cdot \phi) \tau_j \right\}^\top \mathbf{v} \right] \\ &= \partial_v (\mathbf{v} \cdot \phi) \end{aligned} \quad (9.2.9)$$

holds. Here, the following equalities are used:

$$\begin{aligned} \nabla (\mathbf{v} \cdot \phi) \mathbf{v}^\top \mathbf{v} &= \nabla (\mathbf{v} \cdot \phi), \quad \nabla \mathbf{v}^\top \mathbf{v} = \mathbf{0}_{\mathbb{R}^d}, \\ \nabla (\tau_j \cdot \phi) \tau_j^\top \mathbf{v} &= \mathbf{0}_{\mathbb{R}^d}, \quad \mathbf{v} \cdot (\nabla \tau_j^\top \mathbf{v}) = 0. \end{aligned}$$

The fact that $\mathbf{v} \cdot (\nabla \tau_1^\top \mathbf{v}) = 0, \dots, \mathbf{v} \cdot (\nabla \tau_{d-1}^\top \mathbf{v}) = 0$ holds follows from the fact that when $\Omega(\phi)$ is a circle with radius r , such as that in Fig. 9.7, at $x = (0, r)^T$, we have

$$\begin{aligned} \mathbf{v} \cdot (\nabla \tau_j^\top \mathbf{v}) &= (v_1 \ v_2) \begin{pmatrix} \partial \tau_1 / \partial x_1 & \partial \tau_2 / \partial x_1 \\ \partial \tau_1 / \partial x_2 & \partial \tau_2 / \partial x_2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \\ &= (0 \ 1) \begin{pmatrix} 0 & -1/r \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = 0. \end{aligned}$$

Even in a case when $\Omega(\phi)$ is a three-dimensional domain, a similar relationship holds.

Here if Eqs. (9.2.8) and (9.2.9) are substituted into Eq. (9.2.6), then one obtains Eq. (9.2.7). \square

Moreover, the following formula can be obtained with respect to the shape derivative of a normal.

Proposition 9.2.5 (Derivative of the Normal) *For ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, let $\partial\Omega(\phi)$ be piecewise $H^2 \cap C^{0,1}$. In this case, we have the identity*

$$v'(\phi)[\varphi] = -(\nabla\varphi^\top)v + \{v \cdot (\nabla\varphi^\top v)\}v$$

with respect to an arbitrary $\varphi \in Y$. Moreover, $v'(\phi)[\varphi]$ belongs to $C(B; \mathcal{L}(Y; L^\infty(\partial\Omega(\phi); \mathbb{R}^d)))$. \square

Proof The outward unit normal on $\Omega(\phi + \varphi)$ can be expressed as Eq. (9.2.5); that is

$$v(\phi + \varphi) = \frac{\mathbf{F}^{-\top}(\varphi)v}{\|\mathbf{F}^{-\top}(\varphi)v\|_{\mathbb{R}^d}} = \frac{\mathbf{h}(\varphi)}{\|\mathbf{h}(\varphi)\|_{\mathbb{R}^d}}.$$

In this case, we have

$$\begin{aligned} v'(\varphi_0)[\varphi] &= \frac{1}{\|\mathbf{h}(\varphi_0)\|_{\mathbb{R}^d}^2} \left\{ \mathbf{h}'(\varphi_0)[\varphi] \|\mathbf{h}(\varphi_0)\|_{\mathbb{R}^d} - \frac{\mathbf{h}(\varphi_0)^\top (\mathbf{h}'(\varphi_0)[\varphi]) \mathbf{h}(\varphi_0)}{\|\mathbf{h}(\varphi_0)\|_{\mathbb{R}^d}} \right\} \\ &= -(\nabla\varphi^\top)v + \left[v \cdot \{(\nabla\varphi^\top)v\} \right]v. \end{aligned} \quad \square$$

9.3 Shape Derivatives of Functionals

Let us use the results in Sect. 9.2 to obtain the formulae of the shape derivatives of domain and boundary integrals over a moving domain. In this case, one has to be cautious with the two types of formulae of the shape derivatives of domain and boundary integrals: the first one using the shape derivative of a function and the second one using the partial shape derivative of a function.

9.3.1 Formulae Using Shape Derivative of a Function

Firstly, let us consider finding the formulae using the shape derivative u' of a function u . From Definition 9.1.1, the following proposition holds. Here, we assume

that $u(\phi)$ denotes a function u when ϕ , and $f(\phi, u(\phi))$ denotes a functional f when ϕ and $u(\phi)$. Furthermore, write $u'(\phi)[\varphi]$ based on Definition 9.1.1 as u' .

Proposition 9.3.1 (Derivative of Domain Integral of u Using u') *For all ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, suppose $u \in C_{S'}^1(B; L^2(D; \mathbb{R}))$. For an arbitrary $\varphi \in Y$, we set*

$$f(\phi + \varphi, u(\phi + \varphi)) = \int_{\Omega(\phi + \varphi)} u(\phi + \varphi) \, dz.$$

Then,

$$f'(\phi, u)[\varphi] = \int_{\Omega(\phi)} (u' + u \nabla \cdot \varphi) \, dx \quad (9.3.1)$$

holds. Moreover, $f'(\phi, u)[\varphi]$ also belongs to $C(B; \mathcal{L}(X; \mathbb{R}))$. \square

Proof If the domain $\Omega(\phi + \varphi)$ of f is pulled back to $\Omega(\phi)$, we get

$$f(\phi + \varphi, u(\phi + \varphi)) = \int_{\Omega(\phi)} u(\phi + \varphi)(x + \varphi(x)) \omega(\varphi)(x) \, dx.$$

If Definition 9.1.1 is used, one obtains

$$f'(\phi, u(\phi))[\varphi] = \int_{\Omega(\phi)} (u'(\phi)[\varphi] \omega(\varphi_0) + u(\phi) \omega'(\varphi_0)[\varphi]) \, dx.$$

Using Proposition 9.2.1, the desired result follows. \square

Next, let us think about the domain integral when a derivative of a function is the integrand. Firstly, let us focus on the following result. Below we write that the point x on domain D to which $\Omega(\phi)$ is extended moves to $z = x + \varphi(x) = (i + \varphi)(x)$ with respect to an arbitrary $\varphi \in Y$. Moreover, ∇_z represents $\partial(\cdot)/\partial z$.

Proposition 9.3.2 (Pullback of Derivative) *For all ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, let $u \in C(B; H^1(D; \mathbb{R}))$. Suppose*

$$u(\phi + \varphi)(z) = u(\phi)((i + \varphi)^{-1}(z)) = u(\phi)(x) \quad (9.3.2)$$

holds with respect to an arbitrary $\varphi \in Y$. In this case, we have

$$\nabla_z u(\phi + \varphi)(z) = F^{-\top}(\varphi) \nabla u(\phi)(x) \in L^1(D; \mathbb{R}^d).$$

Moreover, $F^{-\top}(\varphi) \nabla u(\phi)$ belongs to $C(B; \mathcal{L}(X; L^1(D; \mathbb{R}^d)))$. \square

Proof The chain rule of derivatives gives

$$\frac{\partial u(\phi + \varphi)}{\partial z}(z) = \frac{\partial \mathbf{x}^\top}{\partial z} \frac{\partial u(\phi)}{\partial \mathbf{x}}(\mathbf{x}) = \left(\frac{\partial z}{\partial \mathbf{x}^\top} \right)^{-\top} \frac{\partial u(\phi)}{\partial \mathbf{x}}(\mathbf{x}). \quad \square$$

Here, if the derivative of a function is included in the integrand of a domain integral, the following formula is obtained [97, 98, 129].

Proposition 9.3.3 (Derivative of Domain Integral of ∇u Using u') *In a neighborhood $B \subset Y$ of $\phi \in \mathcal{D}^\circ$, suppose $u \in C_{S'}^1(B; H^1(D; \mathbb{R}))$. For an arbitrary $\varphi \in Y$, let*

$$f(\phi + \varphi, \nabla_z u(\phi + \varphi)) = \int_{\Omega(\phi + \varphi)} \nabla_z u(\phi + \varphi) \, dz.$$

In this case, the shape derivative of f becomes

$$f'(\phi, \nabla u)[\varphi] = \int_{\Omega(\phi)} \left\{ \nabla u' - (\nabla \varphi^\top) \nabla u + (\nabla \cdot \varphi) \nabla u \right\} \, dx. \quad (9.3.3)$$

Moreover, $f'(\phi, \nabla u)[\varphi]$ also belongs to $C(B; \mathcal{L}(X; \mathbb{R}))$. \square

Proof In Proposition 9.3.2, if we assume Eq. (9.3.2), then we obtain

$$\begin{aligned} f(\phi + \varphi, \nabla_z u(\phi + \varphi)) \\ = \int_{\Omega(\phi + \varphi)} & \left[\nabla_z u(\phi + \varphi)(z)|_* \right. \\ & \left. + \nabla_z \left\{ u(\phi + \varphi)(z) - u(\phi) \left((i + \varphi)^{-1}(z) \right) \right\} \right] \, dz \\ = \int_{\Omega(\phi)} & \left\{ \mathbf{F}^{-\top}(\varphi) \nabla u(\phi)(\mathbf{x}) + u_{\mathbf{x} + \varphi(\mathbf{x})}(\phi + \varphi)(\mathbf{x} + \varphi(\mathbf{x})) \right. \\ & \left. - u_{\mathbf{x} + \varphi(\mathbf{x})}(\phi)(\mathbf{x}) \right\} \omega(\varphi) \, dx. \end{aligned}$$

Here $\nabla_z u(\phi + \varphi)(z)|_*$ is taken to be $\nabla_z u(\phi + \varphi)(z)$ in view of Eq. (9.3.2). From the definition of the shape derivative of f (Definition 9.1.4) and the definition of $u'(\phi)[\varphi]$ (Definition 9.1.1), we obtain

$$\begin{aligned} f'(\phi, \nabla u(\phi))[\varphi] &= \int_{\Omega(\phi)} \left\{ \left(\mathbf{F}^{-\top}(\varphi_0)[\varphi] \nabla u(\phi) + \nabla u'(\phi)[\varphi] \right) \omega(\varphi_0) \right. \\ &\quad \left. + \mathbf{F}^{-\top}(\varphi_0) \nabla u(\phi) \omega'(\varphi_0)[\varphi] \right\} \, dx. \end{aligned}$$

If Propositions 9.2.1 and 9.2.2 are used in this result, then the conclusion follows. \square

A comparison of Propositions 9.3.1 and 9.3.3 suggests the following. With respect to the terms relating to the shape derivative of the domain measure (term containing $\nabla \cdot \boldsymbol{\varphi}$), since the domain measures are only multiplied by $\nabla \cdot \boldsymbol{\varphi}$, both are treated in the same way. On the other hand, with respect to the terms relating to the integrands, the treatments of the two are different. If the integrand does not contain any differential term, u simply changes to u' , but if there is a derivative, ∇u changes to $\nabla u' - (\nabla \boldsymbol{\varphi}^\top) \nabla u$. If attention is given to this point, the following can be obtained if the integrand is given by a function of u and ∇u .

Proposition 9.3.4 (Derivative of Domain Integral Using u') *For all $\boldsymbol{\phi}$ in a neighborhood $B \subset Y$ of $\boldsymbol{\phi}_0 \in \mathcal{D}^\circ$, let $u \in C_{S'}^1(B; \mathcal{U})$ ($\mathcal{U} = H^2(D; \mathbb{R})$) and $h \in C^1(\mathbb{R} \times \mathbb{R}^d; \mathbb{R})$ be defined as*

$$h(u, \nabla u), h_u(u, \nabla u) \in L^2(D; \mathbb{R}), \quad h_{\nabla u}(u, \nabla u) \in L^2(D; \mathbb{R}^d)$$

with respect to $(u, \nabla u) \in \mathcal{U} \times \mathcal{G}$ ($\mathcal{G} = \{\nabla u \mid u \in \mathcal{U}\}$). Let

$$\begin{aligned} f(\boldsymbol{\phi} + \boldsymbol{\varphi}, u(\boldsymbol{\phi} + \boldsymbol{\varphi}), \nabla_z u(\boldsymbol{\phi} + \boldsymbol{\varphi})) \\ = \int_{\Omega(\boldsymbol{\phi} + \boldsymbol{\varphi})} h(u(\boldsymbol{\phi} + \boldsymbol{\varphi}), \nabla_z u(\boldsymbol{\phi} + \boldsymbol{\varphi})) \, dz. \end{aligned}$$

In this case, the shape derivative of f becomes

$$\begin{aligned} f'(\boldsymbol{\phi}, u, \nabla u)[\boldsymbol{\varphi}] \\ = \int_{\Omega(\boldsymbol{\phi})} \{h_u(u, \nabla u)[u'] + h_{\nabla u}(u, \nabla u)\left[\nabla u' - (\nabla \boldsymbol{\varphi}^\top) \nabla u\right] \\ + h(u, \nabla u) \nabla \cdot \boldsymbol{\varphi}\} \, dx. \end{aligned} \quad (9.3.4)$$

Furthermore, $f'(\boldsymbol{\phi}, u, \nabla u)[\boldsymbol{\varphi}]$ also belongs to $C(B; \mathcal{L}(X; \mathbb{R}))$. \square

The formula obtained in Proposition 9.3.4 becomes a key identity used in seeking the shape derivatives of cost functions in Sect. 9.8.1. From the next section onward, $f(\boldsymbol{\phi}, u, \nabla u)$ will be written as $f(\boldsymbol{\phi}, u)$ and Eq. (9.3.4) will be expressed as

$$f'(\boldsymbol{\phi}, u, \nabla u)[\boldsymbol{\varphi}] = f'(\boldsymbol{\phi}, u)[\boldsymbol{\varphi}, u'] = f_{\boldsymbol{\phi}'}(\boldsymbol{\phi}, u)[\boldsymbol{\varphi}] + f_u(\boldsymbol{\phi}, u)[u']. \quad (9.3.5)$$

Here,

$$\begin{aligned} f_{\boldsymbol{\phi}'}(\boldsymbol{\phi}, u)[\boldsymbol{\varphi}] \\ = \int_{\Omega(\boldsymbol{\phi})} \left\{ h_{\nabla u}(u, \nabla u)\left[-(\nabla \boldsymbol{\varphi}^\top) \nabla u\right] + h(u, \nabla u) \nabla \cdot \boldsymbol{\varphi} \right\} \, dx, \end{aligned} \quad (9.3.6)$$

$$f_u(\boldsymbol{\phi}, u)[u'] = \int_{\Omega(\boldsymbol{\phi})} \{h_u(u, \nabla u)[u'] + h_{\nabla u}(u, \nabla u)[\nabla u']\} \, dx. \quad (9.3.7)$$

In the expression of Eq. (9.3.5), all the terms are divided into the linear forms of φ and u' . This formulation will be used when we calculate the shape derivative of the Lagrange function with respect to each cost function. In this situation, we will obtain the shape derivative of each cost function from the linear form of φ and the weak form of adjoint problem from linear form of u' . In Eq. (9.3.5), the subscript $(\cdot)_{\phi'}$ is used to distinguish the similar partial shape derivative shown in Sect. 9.3.2, where $(\cdot)_{\phi^*}$ will be used.

Regarding the second-order shape derivative of the domain integral, we will only check the formulation for the shape derivative of the function to use it later. Here, we focus only on $f_{\phi'}(\phi, u)[\varphi]$, and will show the formulation of $f_{\phi'\phi'}(\phi, u)[\varphi_1, \varphi_2]$. According to Definition 9.1.4, it could be expressed in two parts and is given by

$$f_{\phi'\phi'}(\phi, u)[\varphi_1, \varphi_2] = (f_{\phi'})_{\phi'}(\phi, u)[\varphi_1, \varphi_2] + \langle g(\phi, u), t(\varphi_1, \varphi_2) \rangle, \quad (9.3.8)$$

where

$$(f_{\phi'})_{\phi'}(\phi, u)[\varphi_1, \varphi_2] = \lim_{\|\varphi_2\|_X \rightarrow 0} \frac{1}{\|\varphi_2\|_X} (\langle g(\phi + \varphi_2, u), \varphi_1 \rangle - \langle g(\phi, u), \varphi_1 \rangle), \quad (9.3.9)$$

$$\langle g(\phi, u), t(\varphi_1, \varphi_2) \rangle = \lim_{\|\varphi_2\|_X \rightarrow 0} \frac{1}{\|\varphi_2\|_X} \left\langle g(\phi + \varphi_2, u), \varphi_1 \circ (i + \varphi_2)^{-1} - \varphi_1 \right\rangle. \quad (9.3.10)$$

Equation (9.3.9) represents the derivative of $\langle g(\phi + \varphi_2, u), \varphi_1 \rangle$ with respect to a variation of φ_2 fixing φ_1 . On the other hand, Eq. (9.3.10) is the element to correct the variation of φ_1 by φ_2 using the inverse mapping of $i + \varphi_2$. The calculation of only the term of $\varphi_1 \circ (i + \varphi_2)^{-1} - \varphi_1$ yields Eq. (9.1.9). However, $f_{\phi'}(\phi, u)[\varphi]$ in Eq. (9.3.5) uses $\nabla\varphi^\top$ and $\nabla \cdot \varphi$. Then, we need another formulation shown in the following.

According to the explanation given after Eq. (9.1.9), with respect to $-\nabla\varphi^\top$ in Eq. (9.3.6), we replace φ by φ_1 and add the variation $-\varphi_2$ which is a linearization of the inverse mapping of $i + \varphi_2$. By this variation, $-\nabla\varphi^\top$ becomes $(\nabla\varphi_2^\top - \nabla \cdot \varphi_2) \nabla\varphi_1^\top$ using Proposition 9.3.3 in which φ is changed by $-\varphi_2$, and u is replaced by φ_1^\top . Moreover, we apply the same variation to $\nabla \cdot \varphi$ in Eq. (9.3.6). Using Proposition 9.3.3 in which φ is changed by $-\varphi_2$, and u is replaced by

$\cdot \varphi_1$, $\nabla \cdot \varphi$ becomes $\nabla \varphi_2^\top \cdot (\nabla \varphi_1^\top)^\top - \nabla \cdot \varphi_2 \nabla \cdot \varphi_1$. Using these relations, we have

$$\begin{aligned} & \langle g(\phi, u), t(\varphi_1, \varphi_2) \rangle \\ &= \int_{\Omega(\phi)} \left\{ h_{\nabla u}(u, \nabla u) \left[(\nabla \varphi_2^\top - \nabla \cdot \varphi_2) \nabla \varphi_1^\top \nabla u \right] \right. \\ & \quad \left. + h(u, \nabla u) \left(\nabla \varphi_2^\top \cdot (\nabla \varphi_1^\top)^\top - \nabla \cdot \varphi_2 \nabla \cdot \varphi_1 \right) \right\} dx. \end{aligned} \quad (9.3.11)$$

In calculating the second-order derivatives of cost functions, one has to pay attention to the term in Eq. (9.3.11) added to the expression given in Eq. (9.3.9).

Next, let us think about the case when the functional is given by a boundary integral. Suppose $\Gamma(\phi)$ is a partial set of $\partial\Omega(\phi)$ (allowing $\Gamma(\phi) = \partial\Omega(\phi)$). Moreover, let $\Theta(\phi)$ be corner points (when $d = 2$) or edges (when $d = 3$) on $\partial\Omega(\phi)$ (Fig. 9.3). Also, let τ be a tangent of $\Gamma(\phi)$ (when $d = 2$) or tangent of $\Gamma(\phi)$ and outward normal of $\partial\Gamma(\phi)$ (when $d = 3$). Note that τ at $\Theta(\phi)$ exists on both sides of $\Theta(\phi)$ as shown in Fig. 9.3. Lastly, let $d\zeta$ express the measure of $\partial\Gamma(\phi) \cup \Theta(\phi)$.

Proposition 9.3.5 (Derivative of Boundary Integral of u Using u') *For all ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, let $u \in C_{S'}^1(B; H^1(D; \mathbb{R}))$ and $\Gamma(\phi)$ be piecewise $H^2 \cap C^{0,1}$. For an arbitrary $\varphi \in Y$, let*

$$f(\phi + \varphi, u(\phi + \varphi)) = \int_{\Gamma(\phi+\varphi)} u(\phi + \varphi) d\zeta.$$

Then, the shape derivative of f becomes

$$f'(\phi, u)[\varphi] = \int_{\Gamma(\phi)} \{u' + u(\nabla \cdot \varphi)_\tau\} d\gamma,$$

where $(\nabla \cdot \varphi)_\tau$ follows Eq. (9.2.6). Furthermore, if $\Gamma(\phi)$ is piecewise $H^3 \cap C^{1,1}$,

$$f'(\phi, u)[\varphi] = \int_{\Gamma(\phi)} (u' + \kappa u \mathbf{v} \cdot \varphi - \nabla_\tau u \cdot \varphi_\tau) d\gamma + \int_{\partial\Gamma(\phi) \cup \Theta(\phi)} u \tau \cdot \varphi d\zeta$$

holds, where $\nabla_\tau(\cdot) = (\tau_j(\phi) \cdot \nabla)_{j \in \{1, \dots, d-1\}}(\cdot) \in \mathbb{R}^{d-1}$ and $\varphi_\tau = (\tau_j(\phi) \cdot \varphi)_{j \in \{1, \dots, d-1\}} \in \mathbb{R}^{d-1}$. Moreover, $f'(\phi, u)[\varphi]$ also belongs to $C(B; \mathcal{L}(X; \mathbb{R}))$. \square

Proof If the integral domain $\Gamma(\phi + \varphi)$ of f is pulled back to $\Gamma(\phi)$, we get

$$f(\phi + \varphi, u(\phi + \varphi)) = \int_{\Gamma(\phi)} u(\phi + \varphi)(x + \varphi(x)) \varpi(\varphi) d\gamma.$$

From the definition of the shape derivative of f (Definition 9.1.4) and the definition of $u'(\phi)[\varphi]$ (Definition 9.1.1), we obtain

$$f'(\phi, u(\phi))[\varphi] = \int_{\Gamma(\phi)} \{u'(\phi)[\varphi] \varpi(\varphi_0) + u(\phi) \varpi'(\varphi_0)[\varphi]\} d\gamma.$$

If Proposition 9.2.4 is applied, the first part of the proposition can be obtained. Furthermore, if $\Gamma(\phi)$ is piecewise $H^3 \cap C^{1,1}$, and if the Gauss–Green theorem (Theorem A.8.2) is applied to $\int_{\Gamma(\phi)} u(\phi) \nabla_\tau \cdot \varphi_\tau d\gamma$, the remaining part is established. \square

Furthermore, if the integrand of the boundary integral is a derivative in the direction of the normal, we get the following.

Proposition 9.3.6 (Derivative of Boundary Integral of $\partial_\nu u$ Using u') *For all ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, let $u \in C_{S'}^1(B; H^2(D; \mathbb{R}))$ and $\Gamma(\phi)$ be piecewise $H^2 \cap C^{0,1}$. For an arbitrary $\varphi \in Y$, let*

$$f(\phi + \varphi, \partial_\mu u(\phi + \varphi)) = \int_{\Gamma(\phi + \varphi)} \partial_\mu u(\phi + \varphi) d\zeta.$$

In this case, the shape derivative of f becomes

$$f'(\phi, \partial_\nu u)[\varphi] = \int_{\Gamma(\phi)} \{\partial_\nu u' + w(\varphi, u) + \partial_\nu u (\nabla \cdot \varphi)_\tau\} d\gamma,$$

where

$$w(\varphi, u) = \left[\left\{ \mathbf{v} \cdot (\nabla \varphi^\top \mathbf{v}) \right\} \mathbf{v} - \left\{ (\nabla \varphi^\top + (\nabla \varphi^\top)^\top) \mathbf{v} \right\} \cdot \nabla u \right], \quad (9.3.12)$$

and $(\nabla \cdot \varphi)_\tau$ follows Eq. (9.2.6). Moreover, if $\Gamma(\phi)$ is piecewise $H^3 \cap C^{1,1}$, the identity

$$\begin{aligned} f'(\phi, \partial_\nu u)[\varphi] &= \int_{\Gamma(\phi)} \{\partial_\nu u' + w(\varphi, u) + \kappa \partial_\nu u \mathbf{v} \cdot \varphi - \nabla_\tau (\partial_\nu u) \cdot \varphi_\tau\} d\gamma \\ &\quad + \int_{\partial\Gamma(\phi) \cup \Theta(\phi)} \partial_\nu u \boldsymbol{\tau} \cdot \varphi d\zeta \end{aligned} \quad (9.3.13)$$

holds. Moreover, $f'(\phi, \partial_\nu u)[\varphi]$ belongs to $C(B; \mathcal{L}(Y; \mathbb{R}))$. \square

Proof If we assume Eq. (9.3.2) to hold in Proposition 9.3.2, then we get

$$\begin{aligned}
& f(\phi + \varphi, \partial_\mu u(\phi + \varphi)) \\
&= \int_{\Gamma(\phi + \varphi)} \left[\nabla_z u(\phi + \varphi)(z)|_* \right. \\
&\quad \left. + \nabla_z \left\{ u(\phi + \varphi)(z) - u(\phi) \left((i + \varphi)^{-1}(z) \right) \right\} \right] \cdot \mathbf{v}(\phi + \varphi)(z) d\zeta \\
&= \int_{\Gamma(\phi)} \left\{ \left(\mathbf{F}^{-\top}(\varphi) \nabla u(\phi) \right) \cdot (\mathbf{v} + \mathbf{v}'(\phi)[\varphi] + o(\|\varphi\|_X)) \right. \\
&\quad \left. + \partial_\mu u(\phi + \varphi)(x + \varphi(x)) - \partial_\mu u(x) \right\} \varpi(\varphi) d\gamma,
\end{aligned}$$

where $\nabla_z u(\phi + \varphi)(z)|_*$ equates to $\nabla_z u(\phi + \varphi)(z)$ under the assumption of Eq. (9.3.2). From the definition of the shape derivative of f (Definition 9.1.4) and the definition of $u'(\phi)[\varphi]$ (Definition 9.1.1), we get

$$\begin{aligned}
f'(\phi, \partial_\nu u(\phi))[\varphi] &= \int_{\Gamma(\phi)} \left[\left\{ \left(\mathbf{F}^{-\top}(\varphi_0)[\varphi] \nabla u \right) \cdot \mathbf{v} + \partial_\nu u'(\phi)[\varphi] \right. \right. \\
&\quad \left. \left. + \left(\mathbf{F}^{-\top}(\varphi_0) \nabla u(\phi) \right) \cdot \mathbf{v}'(\phi)[\varphi] \right\} \varpi(\varphi_0) \right. \\
&\quad \left. + \mathbf{F}^{-\top}(\varphi_0) \partial_\nu u(\phi) \varpi'(\varphi_0)[\varphi] \right] d\gamma.
\end{aligned}$$

Using Propositions 9.2.2, 9.2.4 and 9.2.5, we have

$$\begin{aligned}
f'(\phi, \partial_\nu u(\phi))[\varphi] &= \int_{\Gamma(\phi)} \left[- \left\{ \left(\nabla \varphi^\top \right) \nabla u(\phi) \right\} \cdot \mathbf{v} + \partial_\nu u'(\phi)[\varphi] \right. \\
&\quad \left. + \left[- \left(\nabla \varphi^\top \right) \mathbf{v} + \left\{ \mathbf{v} \cdot \left(\nabla \varphi^\top \right) \mathbf{v} \right\} \mathbf{v} \right] \cdot \nabla u(\phi) \right. \\
&\quad \left. + \partial_\nu u(\phi) \left\{ \nabla \cdot \varphi - \mathbf{v} \cdot \left(\left(\nabla \varphi^\top \right) \mathbf{v} \right) \right\} \right] d\gamma.
\end{aligned}$$

From this, the first part of the proposition can be obtained. The remaining part can be obtained in a similar way to the proof of Proposition 9.3.5. \square

When the integrand of a boundary integral is given by the function of u and $\partial_\nu u$, if the chain rule for derivatives is used in the proof of Propositions 9.3.5 and 9.3.6, the following results can be obtained.

Proposition 9.3.7 (Derivative of Boundary Integral Using u') *For all ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, let $u \in C_{S'}^1(B; \mathcal{U})$ ($\mathcal{U} = H^2(D; \mathbb{R})$), and $h \in C^1(\mathbb{R} \times \mathbb{R}; \mathbb{R})$ be defined as*

$$h(u, \partial_\nu u) \in H^2(D; \mathbb{R}), \quad h_u(u, u), h_{\partial_\nu u}(u, \nabla u) \in H^1(D; \mathbb{R}^d)$$

with respect to $(u, \partial_v u) \in \mathcal{U} \times \mathcal{G}_{\Gamma(\phi)}$ ($\mathcal{G}_{\Gamma(\phi)} = \{ \partial_v u|_{\Gamma(\phi)} \mid u \in \mathcal{U} \}$) and $\Gamma(\phi)$ be piecewise $H^2 \cap C^{0,1}$. For an arbitrary $\varphi \in Y$, let

$$\begin{aligned} f(\phi + \varphi, u(\phi + \varphi), \partial_\mu u(\phi + \varphi)) \\ = \int_{\Gamma(\phi+\varphi)} h(u(\phi + \varphi), \partial_\mu u(\phi + \varphi)) d\zeta. \end{aligned}$$

In this case, the shape derivative of f becomes

$$\begin{aligned} f'(\phi, u, \partial_v u)[\varphi] \\ = \int_{\Gamma(\phi)} \{ h_u(u, \partial_v u)[u'] + h_{\partial_v u}(u, \partial_v u)[\partial_v u' + w(\varphi, u)] \\ + h(u, \partial_v u)(\nabla \cdot \varphi)_\tau \} d\gamma. \end{aligned}$$

Here, $w(\varphi, u)$ and $(\nabla \cdot \varphi)_\tau$ are given by Eqs. (9.3.12) and (9.2.6), respectively. Furthermore, if $\Gamma(\phi)$ is piecewise $H^3 \cap C^{1,1}$, we have

$$\begin{aligned} f'(\phi, u, \partial_v u)[\varphi] \\ = \int_{\Gamma(\phi)} \{ h_u(u, \partial_v u)[u'] + h_{\partial_v u}(u, \partial_v u)[\partial_v u' + w(\varphi, u)] \\ + \kappa h(u, \partial_v u) \mathbf{v} \cdot \varphi - \nabla_\tau h(u, \partial_v u) \cdot \varphi_\tau \} d\gamma \\ + \int_{\partial\Gamma(\phi) \cup \Theta(\phi)} h(u, \partial_v u) \boldsymbol{\tau} \cdot \varphi d\zeta. \end{aligned} \quad (9.3.14)$$

Moreover, $f'(\phi, u, \partial_v u)[\varphi]$ belongs to $C(B; \mathcal{L}(Y; \mathbb{R}))$. □

In Propositions 9.3.6 and 9.3.7, we remark the following.

Remark 9.3.8 (Derivative of Boundary Integral of $\partial_v u$ Using u') For a boundary integral that included the derivative of a function, $w(\varphi, u)$ of Eq. (9.3.12) was contained in the shape derivatives of the boundary integral (Eqs. (9.3.13) and (9.3.14)). For that reason, we had $f'(\phi, u, \partial_v u)[\cdot] \in \mathcal{L}(Y; \mathbb{R})$ ($\notin \mathcal{L}(X; \mathbb{R})$). As shown in Sect. 9.1.3, the shape derivatives were defined as bounded linear operators with respect to an arbitrary $\varphi \in X$. Hence in future discussions, when defining the cost functions, the shape derivatives of cost functions must be constructed so that $w(\varphi, u)$ is not left in there. In actual fact, if the cost function is defined as Eq. (9.6.1), the desired results can be obtained. □

The formula obtained in Proposition 9.3.7 is the key identity for obtaining the shape derivative of the cost function in Sect. 9.8.1. From the next section onward, we will write $f(\phi, u, \partial_v u)$ as $f(\phi, u)$, and Eq. (9.3.14) as

$$f'(\phi, u, \partial_v u)[\varphi] = f'(\phi, u)[\varphi, u'] = f_{\phi'}(\phi, u)[\varphi] + f_u(\phi, u)[u']. \quad (9.3.15)$$

Here stands

$$\begin{aligned} f_{\phi'}(\phi, u)[\varphi] &= \int_{\Gamma(\phi)} \{h_{\partial_v u}(u, \partial_v u)[w(\varphi, u)] + h(u, \partial_v u)(\nabla \cdot \varphi)_\tau\} d\gamma, \\ f_u(\phi, u)[u'] &= \int_{\Gamma(\phi)} (h_u(u, \partial_v u)[u'] + h_{\partial_v u}(u, \partial_v u)[\partial_v u']) d\gamma. \end{aligned}$$

Furthermore, if $\Gamma(\phi)$ is piecewise $H^3 \cap C^{1,1}$, we have

$$\begin{aligned} f_{\phi'}(\phi, u)[\varphi] &= \int_{\Gamma(\phi)} \left(h_{\partial_v u}(u, \partial_v u)[w(\varphi, u)] \right. \\ &\quad \left. + \kappa h(u, \partial_v u) \mathbf{v} \cdot \varphi - \nabla_\tau h(u, \partial_v u) \cdot \varphi_\tau \right) d\gamma \\ &\quad + \int_{\partial\Gamma(\phi) \cup \Theta(\phi)} h(u, \partial_v u) \boldsymbol{\tau} \cdot \varphi d\zeta. \end{aligned}$$

9.3.2 Formulae Using Partial Shape Derivative of a Function

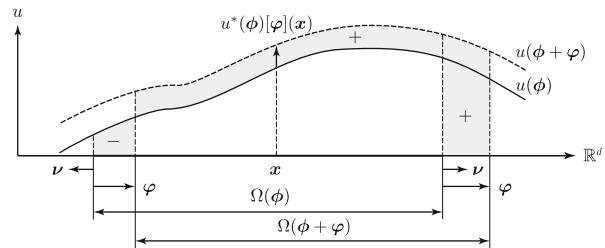
Next, let us use the partial shape derivative u^* of the function u (Definition 9.1.3) to obtain the formulae for seeking the shape derivatives of domain and boundary integrals. Again, we express a function and a functional as $u(\phi + \varphi)$ and $f(\phi + \varphi, u(\phi + \varphi))$, respectively, when $\phi + \varphi$, and simply by u and $f(\phi, u)$ when ϕ . Furthermore, we write $u^*(\phi)[\varphi]$ in Definition 9.1.3 as u^* .

Firstly, in view of Proposition 9.3.1, the following result holds.

Proposition 9.3.9 (Derivative of Domain Integral of u Using u^*) *For all ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, let $u \in C_{S^*}^1(B; H^1(D; \mathbb{R}))$. For an arbitrary $\varphi \in Y$, let*

$$f(\phi + \varphi, u(\phi + \varphi)) = \int_{\Omega(\phi + \varphi)} u(\phi + \varphi) dz.$$

Fig. 9.8 Shape derivative of a domain integral when the partial shape derivative u^* of a function is used



In this case, the shape derivative of f becomes

$$f'(\phi, u)[\varphi] = \int_{\Omega(\phi)} u^* dx + \int_{\partial\Omega(\phi)} u \mathbf{v} \cdot \varphi d\gamma. \quad (9.3.16)$$

Moreover, $f'(\phi, u)[\varphi]$ also belongs to $C(B; \mathcal{L}(X; \mathbb{R}))$.

Proof The proposition is easily proved by substituting $u'(\phi)[\varphi]$ in Eq. (9.1.4) of Proposition 9.3.1 together with the Gauss–Green theorem (Theorem A.8.2). \square

Figure 9.8 shows the areas corresponding to each integral on the right-hand side of Eq. (9.3.16). The first term on the right-hand side corresponds to the shaded area in $\Omega(\phi) \cap \Omega(\phi + \varphi)$, while the second term corresponds to the shaded areas on the left and right sides of the figure. Here, it should be noted that since the area on the right side has the outward unit normal \mathbf{v} pointing to the right, $\mathbf{v} \cdot \varphi > 0$, and since the area on the left side has \mathbf{v} pointing to the left, then $\mathbf{v} \cdot \varphi < 0$.

Moreover, the formula corresponding to Proposition 9.3.3 which represents the shape derivative of the domain integral with a differential term as integrand can be obtained by viewing $\nabla u \in C_{S^*}^1(B; H^1(D; \mathbb{R}^d))$ as u in Proposition 9.3.9. In this case, from the fact that $(\nabla u)^*(\phi)[\varphi] = \nabla u^*(\phi)[\varphi]$ can be established based on Definition 9.1.3, if $\nabla u^*(\phi)[\varphi]$ is written as ∇u^* , then we have

$$f'(\phi, \nabla u)[\varphi] = \int_{\Omega(\phi)} \nabla u^* dx + \int_{\partial\Omega(\phi)} (\nu \cdot \varphi) \nabla u dy. \quad (9.3.17)$$

Furthermore, Eq. (9.3.17) can be written as

$$\begin{aligned} f'(\boldsymbol{\phi}, \nabla u)[\boldsymbol{\varphi}] &= \int_{\Omega(\boldsymbol{\phi})} \left[\nabla u^* + \left\{ \nabla^\top \left(\nabla u \boldsymbol{\varphi}^\top \right)^\top \right\}^\top \right] \mathrm{d}x \\ &= \int_{\Omega(\boldsymbol{\phi})} (\nabla u^* + \nabla \cdot \boldsymbol{\varphi} \nabla u + \Delta u \boldsymbol{\varphi}) \mathrm{d}x \end{aligned} \quad (9.3.18)$$

using the Gauss–Green theorem. Hence, if it is compared with the results of Proposition 9.3.3, we get the identity

$$\nabla u'(\phi)[\varphi] = \nabla u^*(\phi)[\varphi] + \left(\nabla \varphi^\top \right) \nabla u(\phi) + \Delta u(\phi) \varphi. \quad (9.3.19)$$

Equation (9.3.19) can be obtained also from

$$\nabla u'(\phi)[\varphi] = \nabla u^*(\phi)[\varphi] + \nabla(\nabla u(\phi) \cdot \varphi)$$

by using Eq. (9.1.4).

If the integrand is given by a function of u and ∇u , then by using the chain rule for derivatives on Proposition 9.3.9, the following result can be obtained.

Proposition 9.3.10 (Derivative of Domain Integral Using u^*) *For all ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, let $u \in C_{S^*}^1(B; \mathcal{U})$ ($\mathcal{U} = H^2(D; \mathbb{R})$) and $h \in C^1(\mathbb{R} \times \mathbb{R}^d; \mathbb{R})$ be defined as*

$$h(u, \nabla u) \in H^1(D; \mathbb{R}), \quad h_u(u, \nabla u), h_{\nabla u}(u, \nabla u) \in L^2(D; \mathbb{R}^d)$$

with respect to $(u, \nabla u) \in \mathcal{U} \times \mathcal{G}$ ($\mathcal{G} = \{\nabla u \mid u \in \mathcal{U}\}$). For an arbitrary $\varphi \in Y$, let

$$\begin{aligned} f(\phi + \varphi, u(\phi + \varphi), \nabla_z u(\phi + \varphi)) \\ = \int_{\Omega(\phi + \varphi)} h(u(\phi + \varphi), \nabla_z u(\phi + \varphi)) \, dz. \end{aligned}$$

Then, the shape derivative of f in this case becomes

$$\begin{aligned} f'(\phi, u, \nabla u)[\varphi] &= \int_{\Omega(\phi)} \{h_u(u, \nabla u)[u^*] + h_{\nabla u}(u, \nabla u)[\nabla u^*]\} \, dx \\ &\quad + \int_{\partial\Omega(\phi)} h(u, \nabla u) \, v \cdot \varphi \, d\gamma. \end{aligned} \quad (9.3.20)$$

Moreover, $f'(\phi, u, \nabla u)[\varphi]$ also belongs to $C(B; \mathcal{L}(X; \mathbb{R}))$. □

The formula obtained in Proposition 9.3.10 is the key identity for obtaining the shape derivative of the cost function in Sect. 9.8.4. From now on, we will write $f(\phi, u, \nabla u)$ as $f(\phi, u)$ and Eq. (9.3.20) as

$$f'(\phi, u, \nabla u)[\varphi] = f'(\phi, u)[\varphi, u^*] = f_{\phi^*}(\phi, u)[\varphi] + f_u(\phi, u)[u^*]. \quad (9.3.21)$$

Here, stands

$$f_{\phi^*}(\phi, u)[\varphi] = \int_{\partial\Omega(\phi)} h(u, \nabla u) \mathbf{v} \cdot \varphi \, d\gamma,$$

$$f_u(\phi, u)[u^*] = \int_{\Omega(\phi)} \{h_u(u, \nabla u)[u^*] + h_{\nabla u}(u, \nabla u)[\nabla u^*]\} \, dx.$$

If a functional is given by a boundary integral, the following formula is obtained by substituting Eq. (9.1.4) into Proposition 9.3.5.

Proposition 9.3.11 (Derivative of Boundary Integral of u Using u^*) *For all ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, let $u \in C_{S^*}^1(B; H^2(D; \mathbb{R}))$. For an arbitrary $\varphi \in Y$, let*

$$f(\phi + \varphi, u(\phi + \varphi)) = \int_{\Gamma(\phi + \varphi)} u(\phi + \varphi) \, d\zeta.$$

In this case, the shape derivative of f becomes

$$f'(\phi, u)[\varphi] = \int_{\Gamma(\phi)} (u^* + \nabla u \cdot \varphi + u(\nabla \cdot \varphi)_\tau) \, d\gamma, \quad (9.3.22)$$

where $(\nabla \cdot \varphi)_\tau$ obeys Eq. (9.2.6). Furthermore, if $\Gamma(\phi)$ is piecewise $H^3 \cap C^{1,1}$, we have

$$f'(\phi, u)[\varphi] = \int_{\Gamma(\phi)} \{u^* + (\partial_v + \kappa) u \mathbf{v} \cdot \varphi\} \, d\gamma$$

$$+ \int_{\partial\Gamma(\phi) \cup \Theta(\phi)} u \boldsymbol{\tau} \cdot \varphi \, d\zeta. \quad (9.3.23)$$

Moreover, $f'(\phi, u)[\varphi]$ also belongs to $C(B; \mathcal{L}(X; \mathbb{R}))$. □

Moreover, if the integrand of the boundary integral is $\partial_v u$, the following result is obtained.

Proposition 9.3.12 (Derivative of Boundary Integral of $\partial_v u$ Using u^*) *For all ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, let $u \in C_{S^*}^1(B; H^3(D; \mathbb{R}))$. For an arbitrary $\varphi \in Y$, let*

$$f(\phi + \varphi, \partial_\mu u(\phi + \varphi)) = \int_{\Gamma(\phi + \varphi)} \partial_\mu u(\phi + \varphi) \, d\zeta.$$

In this case, the shape derivative of f becomes

$$f'(\phi, \partial_v u)[\varphi] = \int_{\Gamma(\phi)} (\partial_v u^* + \bar{w}(\varphi, u) + \partial_v u(\nabla \cdot \varphi)_\tau) \, d\gamma,$$

where

$$\bar{w}(\boldsymbol{\varphi}, u) = - \left[\sum_{i \in \{1, \dots, d-1\}} \left\{ \boldsymbol{\tau}_i \cdot \left(\nabla \boldsymbol{\varphi}^\top \mathbf{v} \right) \right\} \boldsymbol{\tau}_i \right] \cdot \nabla u + (\mathbf{v} \cdot \boldsymbol{\varphi}) \Delta u, \quad (9.3.24)$$

and $(\nabla \cdot \boldsymbol{\varphi})_\tau$ obeys Eq. (9.2.6). Furthermore, if $\Gamma(\boldsymbol{\phi})$ is piecewise $H^3 \cap C^{1,1}$, we have

$$\begin{aligned} f'(\boldsymbol{\phi}, \partial_\nu u)[\boldsymbol{\varphi}] &= \int_{\Gamma(\boldsymbol{\phi})} \left\{ \partial_\nu u^* + \bar{w}(\boldsymbol{\varphi}, u) + \kappa \partial_\nu u \mathbf{v} \cdot \boldsymbol{\varphi} \right. \\ &\quad \left. - \nabla_\tau (\partial_\nu u) \cdot \boldsymbol{\varphi}_\tau \right\} d\gamma + \int_{\partial\Gamma(\boldsymbol{\phi}) \cup \Theta(\boldsymbol{\phi})} \partial_\nu u \boldsymbol{\tau} \cdot \boldsymbol{\varphi} d\zeta. \end{aligned}$$

Moreover, $f'(\boldsymbol{\phi}, \partial_\nu u)[\boldsymbol{\varphi}]$ belongs to $C(B; \mathcal{L}(Y; \mathbb{R}))$. \square

Proof From Eq. (9.3.19), we have the equation

$$\partial_\nu u'(\boldsymbol{\phi})[\boldsymbol{\varphi}] = \partial_\nu u^*(\boldsymbol{\phi})[\boldsymbol{\varphi}] + \left\{ \left(\nabla \boldsymbol{\varphi}^\top \right)^\top \mathbf{v} \right\} \cdot \nabla u(\boldsymbol{\phi}) + \Delta u(\boldsymbol{\phi}) \mathbf{v} \cdot \boldsymbol{\varphi}. \quad (9.3.25)$$

Substituting the above equation into the result of Proposition 9.3.6, we arrive at the desired result. \square

Here, if the integrand of a boundary integral is given by a function of u and $\partial_\nu u$, the following result can be obtained by using the chain rule for derivatives on Propositions 9.3.11 and 9.3.12.

Proposition 9.3.13 (Derivative of Boundary Integral Using u^*) For all $\boldsymbol{\phi}$ in a neighborhood $B \subset Y$ of $\boldsymbol{\phi}_0 \in \mathcal{D}^\circ$, let $u \in C_{S^*}^1(B; \mathcal{U})$ ($\mathcal{U} = H^3(D; \mathbb{R})$), and $h \in C^1(\mathbb{R} \times \mathbb{R}; \mathbb{R})$ be defined as

$$h(u, \partial_\nu u) \in H^2(D; \mathbb{R}), \quad h_u(u, u), h_{\partial_\nu u}(u, \nabla u) \in H^1(D; \mathbb{R}^d)$$

with respect to $(u, \partial_\nu u) \in \mathcal{U} \times \mathcal{G}_{\Gamma(\boldsymbol{\phi})}$ ($\mathcal{G}_{\Gamma(\boldsymbol{\phi})} = \{ \partial_\nu u|_{\Gamma(\boldsymbol{\phi})} \mid u \in \mathcal{U} \}$). For an arbitrary $\boldsymbol{\varphi} \in Y$, let

$$\begin{aligned} f(\boldsymbol{\phi} + \boldsymbol{\varphi}, u(\boldsymbol{\phi} + \boldsymbol{\varphi}), \partial_\mu u(\boldsymbol{\phi} + \boldsymbol{\varphi})) \\ = \int_{\Gamma(\boldsymbol{\phi} + \boldsymbol{\varphi})} h(u(\boldsymbol{\phi} + \boldsymbol{\varphi}), \partial_\mu u(\boldsymbol{\phi} + \boldsymbol{\varphi})) d\zeta. \end{aligned}$$

In this case, the shape derivative of f becomes

$$\begin{aligned} f'(\phi, u, \partial_v u) [\varphi] &= \int_{\Gamma(\phi)} \{h_u(u, \partial_v u)[u^*] + \nabla h(u, \partial_v u) \cdot \varphi \\ &\quad + h_{\partial_v u}(u, \partial_v u)[\partial_v u^* + \bar{w}(\varphi, u)] + h(u, \partial_v u)(\nabla \cdot \varphi)_\tau\} d\gamma, \end{aligned}$$

where $\bar{w}(\varphi, u)$ and $(\nabla \cdot \varphi)_\tau$ obey Eqs. (9.3.24) and (9.2.6), respectively. Furthermore, if $\Gamma(\phi)$ is piecewise $H^3 \cap C^{1,1}$, we have

$$\begin{aligned} f'(\phi, u, \partial_v u) [\varphi] &= \int_{\Gamma(\phi)} \{h_u(u, \partial_v u)[u^*] + h_{\partial_v u}(u, \partial_v u)[\partial_v u^* + \bar{w}(\varphi, u)] \\ &\quad + (\partial_v + \kappa)h(u, \partial_v u)\mathbf{v} \cdot \varphi\} d\gamma \\ &\quad + \int_{\partial\Gamma(\phi) \cup \Theta(\phi)} h(u, \partial_v u)\boldsymbol{\tau} \cdot \varphi d\zeta. \end{aligned} \tag{9.3.26}$$

Moreover, $f'(\phi, u, \partial_v u) [\varphi]$ belongs to $C(B; \mathcal{L}(Y; \mathbb{R}))$. \square

In Propositions 9.3.12 and 9.3.13, let us recall the similar situation in Remark 9.3.8.

The formula given in Proposition 9.3.13 is the key identity for obtaining the shape derivative of the cost function in Sect. 9.8.4. From the next section onward, by writing $f(\phi, u, \partial_v u)$ as $f(\phi, u)$, Eq. (9.3.26) is expressed as

$$f'(\phi, u, \partial_v u) [\varphi] = f'(\phi, u)[\varphi, u^*] = f_{\phi^*}(\phi, u)[\varphi] + f_u(\phi, u)[u^*], \tag{9.3.27}$$

where

$$\begin{aligned} f_{\phi^*}(\phi, u)[\varphi] &= \int_{\Gamma(\phi)} \{h_u(u, \partial_v u)[\nabla u(\phi) \cdot \varphi] + h_{\partial_v u}(u, \partial_v u)[\bar{w}(\varphi, u)] \\ &\quad + h(u(\phi), \partial_v u(\phi))(\nabla \cdot \varphi)_\tau\} d\gamma, \\ f_u(\phi, u)[u^*] &= \int_{\Gamma(\phi)} (h_u(u, \partial_v u)[u^*] + h_{\partial_v u}(u, \partial_v u)[\partial_v u^*]) d\gamma. \end{aligned}$$

Furthermore, if $\Gamma(\phi)$ is piecewise $H^3 \cap C^{1,1}$ class,

$$\begin{aligned} f_{\phi^*}(\phi, u)[\varphi] &= \int_{\Gamma(\phi)} \left\{ h_{\partial_v u}(u, \partial_v u) [\bar{w}(\phi, u)] \right. \\ &\quad \left. + (\partial_v + \kappa) h(u(\phi), \partial_v u(\phi)) v \cdot \varphi \right\} d\gamma \\ &\quad + \int_{\partial\Gamma(\phi) \cup \Theta(\phi)} h(u(\phi), \partial_v u(\phi)) \tau \cdot \varphi d\varsigma. \end{aligned}$$

9.4 Variation Rules of Functions

In Sect. 9.5, a state determination problem (boundary value problem of partial differential equation) will be defined. In this case, one has to be aware of how the known function behaves with respect to the moving domain. Here, let us define typical variation rules using the results obtained up to the end of Sect. 9.3. Also, in this section, we will fix $\phi_0 \in \mathcal{D}^\circ$ and consider an arbitrary domain variation $\varphi \in Y$ for ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$.

Firstly, we think about the case when the function value moves along with the movement of a point on the domain, as shown in Fig. 9.9. The variation rule for the function in this case is defined as follows.

Definition 9.4.1 (Function Fixed with Material) For all ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, let $u \in C_{S'}^1(B; L^2(D; \mathbb{R}))$, and suppose

$$u'(\phi)[\varphi] = 0$$

with respect to an arbitrary $\varphi \in Y$. Then, u is referred to as a function fixed with material. \square

Moreover, the variation rule for a function not depending on the domain variation such as that in Fig. 9.10 is defined as follows.

Fig. 9.9 The function $u : D \rightarrow \mathbb{R}$ fixed with material

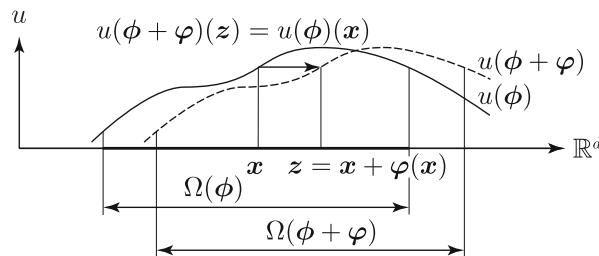


Fig. 9.10 The function $u : D \rightarrow \mathbb{R}$ fixed in space

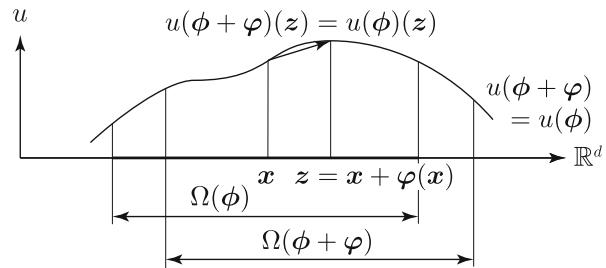
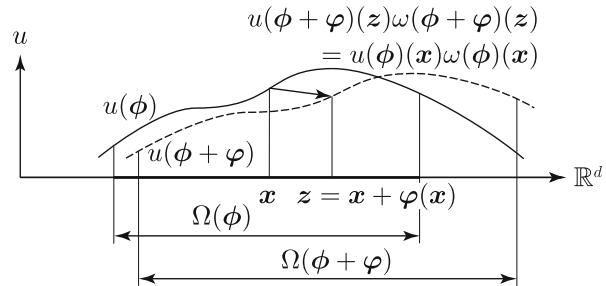


Fig. 9.11 The function $u : D \rightarrow \mathbb{R}$ varying with domain measure



Definition 9.4.2 (Function Fixed in Space) For all ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, let $u \in C_{S^*}^1(B; H^1(D; \mathbb{R}))$, and suppose

$$u'(\phi)[\varphi] - \nabla u(\phi) \cdot \varphi = u^*(\phi)[\varphi] = 0$$

with respect to an arbitrary $\varphi \in Y$. Then, u is referred to as a function fixed in space. \square

Furthermore, consider the case when along with the movement of a point on the domain, its function value changes inversely proportionate to the Jacobian $\omega(\varphi)$ of the domain. Here, the equation

$$\begin{aligned} & u(\phi + \varphi)(x + \varphi(x)) \\ &= \frac{u(\phi)(x)}{\omega(\varphi)(x + \varphi(x))} \\ &= u(\phi)(x) \left(1 - \omega'(\phi_0)[\varphi](x) + o(\|\varphi(x)\|_{\mathbb{R}^d}) \right) \end{aligned} \quad (9.4.1)$$

holds at almost everywhere $x \in D$, see Fig. 9.11 for an illustration. Hence, using Proposition 9.2.1, the variation rule in this case is defined as follows.

Definition 9.4.3 (Function Varying with Domain Measure) For all ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, let $u \in C_{S'}^1(B; L^2(D; \mathbb{R}))$, and suppose

$$u'(\phi)[\varphi] + u(\phi) \nabla \cdot \varphi = 0 \quad (9.4.2)$$

with respect to an arbitrary $\varphi \in Y$. Then, u is called a function varying with domain measure. \square

If Eq. (9.4.2) is substituted into Proposition 9.3.1, then we obtain $f'(\phi, u(\phi))[\varphi] = 0$. Hence, a function varying with a domain measure indicates that the domain integral of the function would be fixed even when the domain varies.

Moreover, if along with the movement of a point on the boundary, its function takes a value inversely proportional to the Jacobian $\varpi(\varphi)$ on the boundary, the equation

$$\begin{aligned} & u(\phi + \varphi)(x + \varphi(x)) \\ &= \frac{u(\phi)(x)}{\varpi(\varphi)(x + \varphi(x))} \\ &= u(\phi)(x) \left(1 - \varpi'(\varphi_0)[\varphi](x) + o(\|\varphi(x)\|_{\mathbb{R}^d}) \right) \end{aligned} \quad (9.4.3)$$

holds. Here, Proposition 9.2.4 is used in order to define the variation rule given in the following definition.

Definition 9.4.4 (Function Varying with Boundary Measure) For all ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^\circ$, let $u \in C_{S'}^1(B; H^1(D; \mathbb{R}))$ and $\partial\Omega(\phi)$ be piecewise $H^2 \cap C^{0,1}$, and suppose

$$u'(\phi)[\varphi] + u(\phi)(\nabla \cdot \varphi)_\tau = 0 \quad (9.4.4)$$

with respect to an arbitrary $\varphi \in Y$ at almost every $x \in \partial\Omega(\phi)$. Then, u is referred to as a function varying with boundary measure. Here, $\nabla_\tau \cdot \varphi$ follows Eq. (9.2.6). \square

If Eq. (9.4.4) is substituted into Proposition 9.3.5, then we obtain $f'(\phi, u(\phi))[\varphi] = 0$. In this case, it indicates the fact that the boundary integral of u remains unchanged.

Let us think about a specific problem using the definition above. Figure 9.12 shows the representative variation patterns when the traction p_0 in a linear elastic

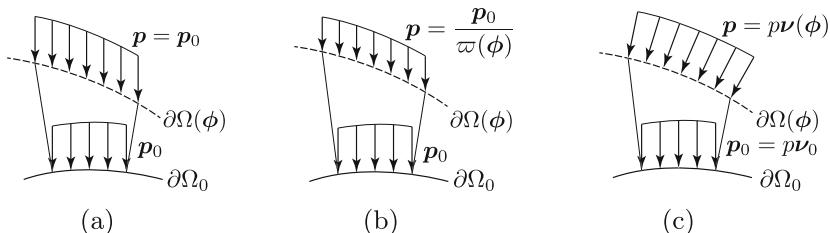


Fig. 9.12 Typical variation patterns of traction p in linear elastic problem. (a) p_0 is constant fixed with material or fixed in space. (b) p_0 is constant varying with boundary measure. (c) p is fixed with space, ν is fixed with material (hydrostatic pressure)

problem moves to p along with the movement of the boundary. Figure 9.12c represents the change in traction on a boundary when the hydrostatic pressure is acting on it. In this book, although the assumption of hydrostatic pressure will not be used directly, in order to use it in future discussion, let us obtain the shape derivative with respect to the boundary integral of hydrostatic pressure.

Proposition 9.4.5 (Derivative of Integral Using Hydrostatic Pressure) *Let $p \in H^2(D; \mathbb{R})$ be a function fixed in space. For ϕ in a neighborhood $B \subset Y$ of $\phi_0 \in \mathcal{D}^o$, let*

$$f(\phi + \varphi, p) = \int_{\Gamma(\phi + \varphi)} p v(\phi + \varphi) d\zeta$$

where $\varphi \in Y$ is arbitrary. In this case, the shape derivative of f becomes

$$f'(\phi, p)[\varphi] = \int_{\Gamma(\phi)} \left\{ (\nabla p \cdot \varphi) v - p (\nabla \varphi^\top) v + p (\nabla \cdot \varphi) v \right\} d\gamma.$$

Moreover, $f'(\phi, p)[\varphi]$ belongs to $C(B; \mathcal{L}(Y; \mathbb{R}))$. □

Proof If $\Gamma(\phi + \varphi)$ is pulled back to $\Gamma(\phi)$, then we have

$$f(\phi + \varphi, p) = \int_{\Gamma(\phi)} p(x + \varphi(x)) v(\phi + \varphi)(x + \varphi(x)) \varpi(\varphi)(x) d\gamma.$$

From the definition of the shape derivative of f ,

$$f'(\phi, p)[\varphi] = \int_{\Gamma(\phi)} \left\{ (p'(\phi)[\varphi] v + p v'(\phi)[\varphi]) \varpi(\varphi_0) + p v \varpi'(\varphi_0)[\varphi] \right\} d\gamma$$

can be obtained. Here, since p is fixed in space (Definition 9.4.2), then the equation $p'(\phi)[\varphi] = \nabla p \cdot \varphi$ holds and if Propositions 9.2.4 and 9.2.5 are used, the desired result then follows. □

9.5 State Determination Problem

Since the definitions and formulas of shape derivatives of functions and functionals have been obtained, let us use them to define a boundary value problem of a partial differential equation which would be a state determination problem. In this chapter, a Poisson problem will be considered first for ease.

In a shape optimization problem of domain variation type, the domains of known functions and the solution function vary along with each other. Let $b_0 : D \rightarrow \mathbb{R}$, $p_{N0} : D \rightarrow \mathbb{R}$, $u_{D0} : D \rightarrow \mathbb{R}$ be known functions over the reference domain Ω_0 , which can then be recovered through a specified variation rule with the functions

$b(\phi) : D \rightarrow \mathbb{R}$, $p_N(\phi) : D \rightarrow \mathbb{R}$, $u_D(\phi) : D \rightarrow \mathbb{R}$ defined over the perturbed domain $\Omega(\phi)$. We shall use their respective variation rules when we eventually deal with computing the shape derivative of an associated cost function.

With respect to the solution function, since it is a function of H^1 class, the Calderón extension theorem (Theorem 4.4.4) can be used to view it as a function defined on D . Hence, we define the real Hilbert space (linear space of state variables in optimal design problem) containing the homogeneous solution (given by $\tilde{u} = u - u_D$ with a known function u_D providing the Dirichlet condition) for the solution of a state determination problem by

$$U(\phi) = \left\{ u \in H^1(D; \mathbb{R}) \mid u = 0 \text{ on } \Gamma_D(\phi) \right\} \quad (9.5.1)$$

with respect to $\phi \in \mathcal{D}$. Furthermore, in order for the domain variation obtained from the gradient method shown later to be in \mathcal{D} of Eq. (9.1.3), the admissible set of state variables for the homogeneous solution \tilde{u} with respect to a state determination problem is taken to be

$$\mathcal{S}(\phi) = U(\phi) \cap W^{2,4}(D; \mathbb{R}). \quad (9.5.2)$$

The regularity which is needed in addition to the condition of $\mathcal{S}(\phi)$ will be specified when required.

The following two types of hypotheses are set with respect to regularity of known functions. When the shape derivatives are sought using formulae based on the shape derivative of a function, the following hypothesis is used later.

Hypothesis 9.5.1 (Known Functions (Shape Derivative)) With respect to the given known functions, in a neighborhood $B \subset Y$ of $\phi \in \mathcal{D}^\circ$, we assume

$$\begin{aligned} b &\in C_{S'}^1\left(B; C^{0,1}(D; \mathbb{R})\right), & p_N &\in C_{S'}^1\left(B; C^{1,1}(D; \mathbb{R})\right), \\ u_D &\in C_{S'}^1\left(B; W^{2,4}(D; \mathbb{R})\right) \end{aligned}$$

and denote their shape derivatives as $(\cdot)'(\phi)[\varphi]$.

On the other hand, the following hypothesis is used when seeking the shape derivatives using the formulae based on the partial shape derivative of a function.

Hypothesis 9.5.2 (Known Functions (Partial Shape Derivative)) With respect to the given known functions, in a neighborhood $B \subset Y$ of $\phi \in \mathcal{D}^\circ$, we assume

$$\begin{aligned} b &\in C_{S^*}^1\left(B; C^{0,1}(D; \mathbb{R})\right), & p_N &\in C_{S^*}^1\left(B; C^{1,1}(D; \mathbb{R})\right), \\ u_D &\in C_{S'}^1\left(B; W^{2,2q_R}(D; \mathbb{R})\right) \end{aligned}$$

where $q_R > d$, and denote their partial shape derivatives as $(\cdot)^*(\phi)[\varphi]$.

The following hypothesis is established with respect to regularity of the boundary.

Hypothesis 9.5.3 (Opening Angle of Corner Point) Let $\Omega(\phi)$ be a two-dimensional domain and consider a corner point on the boundary. When $\Omega(\phi)$ is a three-dimensional domain, we consider a plane which is perpendicular to the corner line on the boundary and the corner point on the boundary in the plane. Let β be the opening angle of the corner point between two boundaries that are a Dirichlet boundary or Neumann boundary,

- (1) if the boundaries are same of the type, assume $\beta < 2\pi/3$,
- (2) if the boundaries are of mixed type, assume $\beta < \pi/3$.

If Hypotheses 9.5.1 and 9.5.3 hold, the fact that u is in \mathcal{S} is shown by Proposition 5.3.1.

Using the hypotheses above, a Poisson problem of domain variation type will be defined as follows. Here, we write $\partial_v = v \cdot \nabla$.

Problem 9.5.4 (Poisson Problem of Domain Variation Type) Let $\phi \in \mathcal{D}$ and $b(\phi)$, $p_N(\phi)$, $u_D(\phi)$ be given. Find $u : \Omega(\phi) \rightarrow \mathbb{R}$ which satisfies

$$\begin{aligned} -\Delta u &= b(\phi) \quad \text{in } \Omega(\phi), \\ \partial_v u &= p_N(\phi) \quad \text{on } \Gamma_p(\phi), \\ \partial_v u &= 0 \quad \text{on } \Gamma_N(\phi) \setminus \bar{\Gamma}_p(\phi), \\ u &= u_D(\phi) \quad \text{on } \Gamma_D(\phi). \end{aligned}$$

□

Here and in what follows, $b(\phi)$ or $u_D(\phi)$ and $U(\phi)$ or $\mathcal{S}(\phi)$, etc. will be written respectively as b or u_D and U or \mathcal{S} , etc.

Problem 9.5.4 will be used as an equality constraint in the shape optimization problem (Problem 9.6.3) of domain variation type shown later. In a later argument, an equality constraint will be replaced with stationary conditions for a Lagrange function. Here, as a preparation for this, we define the Lagrange function of Problem 9.5.4 as

$$\begin{aligned} \mathcal{L}_S(\phi, u, v) &= \int_{\Omega(\phi)} (-\nabla u \cdot \nabla v + bv) \, dx + \int_{\Gamma_p(\phi)} p_N v \, dy \\ &\quad + \int_{\Gamma_D(\phi)} \{(u - u_D) \partial_v v + v \partial_v u\} \, dy, \end{aligned} \quad (9.5.3)$$

where u is not necessarily the solution of Problem 9.5.4 and v is an element of \mathcal{S} introduced as a Lagrange multiplier. In Eq. (9.5.3), the third term on the right-hand side was added in order to make the later discussions easier in a similar way to Eq. (8.2.4) in Chap. 8 defining the Lagrange function with respect to a θ -type Poisson problem. Moreover, in a similar manner to Eq. (7.2.3) defining the Lagrange function with respect to the abstract variational problem in Chap. 7, using $\tilde{u} =$

$u - u_D$, we write

$$\mathcal{L}_S(\phi, u, v) = -a(\phi)(u, v) + l(\phi)(v) = -a(\phi)(\tilde{u}, v) + \hat{l}(\phi)(v), \quad (9.5.4)$$

where

$$a(\phi)(u, v) = \int_{\Omega(\phi)} \nabla u \cdot \nabla v \, dx, \quad (9.5.5)$$

$$l(\phi)(v) = \int_{\Omega(\phi)} bv \, dx + \int_{\Gamma_p(\phi)} p_N v \, d\gamma, \quad (9.5.6)$$

$$\hat{l}(\phi)(v) = l(\phi)(v) + a(\phi)(u_D, v). \quad (9.5.7)$$

When u is the solution to Problem 9.5.4,

$$\mathcal{L}_S(\phi, u, v) = 0$$

holds for all $v \in U$. This equation is equivalent to the weak form of Problem 9.5.4.

Following the notation in Sect. 9.3, $\mathcal{L}_S(\phi, u, v)$ should be written as $\mathcal{L}_S(\phi, u, \nabla u, \partial_\nu u, v, \nabla v, \partial_\nu v)$. However, from now on, it will be written as $\mathcal{L}_S(\phi, u, v)$.

9.6 Shape Optimization Problem of Domain Variation Type

In Sect. 9.5, we saw how the state variable $\tilde{u} = u - u_D \in \mathcal{S}$ is determined as the solution of a state determination problem when a design variable $\phi \in \mathcal{D}$ is given. These variables are used to define a shape optimization problem.

Here, the cost functions are set to

$$\begin{aligned} f_i(\phi, u) &= \int_{\Omega(\phi)} \xi_i(\phi, u, \nabla u) \, dx + \int_{\Gamma_{\eta i}(\phi)} \eta_{Ni}(\phi, u) \, d\gamma \\ &\quad - \int_{\Gamma_D(\phi)} \eta_{Di}(\phi, \partial_\nu u) \, d\gamma - c_i, \end{aligned} \quad (9.6.1)$$

for every $i \in \{0, 1, \dots, m\}$, respectively. Here c_1, \dots, c_m are constants and have to be determined such that there exists some $(\phi, \tilde{u}) \in \mathcal{D} \times \mathcal{S}$ which satisfies $f_i \leq 0$ for all $i \in \{1, \dots, m\}$. Moreover, ξ_i , η_{Ni} and η_{Di} are assumed to be given and satisfy two types of hypotheses as follows. Those hypotheses will be needed to obtain an appropriate regularity in the solution of a adjoint problem (Problem 9.8.1) shown later. To calculate the second-order shape derivatives of cost functions, additional hypotheses are required. However, details of these conditions will be omitted and

we shall only tacitly assume that they were already satisfied to carry out a second-order differentiation of the costs.

The following assumption is used when employing the formulae based on the shape derivative of a function.

Hypothesis 9.6.1 (Cost Functions (Shape Derivative)) With respect to cost function f_i ($i \in \{0, 1, \dots, m\}$) of Eq. (9.6.1), let $\zeta_i \in C^1(\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d; \mathbb{R})$, $\eta_{Ni} \in C^1(\mathbb{R}; \mathbb{R})$, $\eta_{Di} \in C^1(\mathbb{R}; \mathbb{R})$ be functions fixed with material satisfying

$$\begin{aligned} \zeta_i(\phi, u, \nabla u), \zeta_{i\phi'}(\phi, u, \nabla u)[\varphi] &\in H^1 \cap L^\infty(D; \mathbb{R}), \\ \zeta_{iu}(\phi, u, \nabla u)[\hat{u}] &\in L^4(D; \mathbb{R}), \quad \zeta_{i(\nabla u)^\top}(\phi, u, \nabla u)[\nabla \hat{u}] \in W^{1,4}(D; \mathbb{R}^d), \\ \eta_{Ni}(\phi, u), \eta_{Ni\phi'}(\phi, u)[\varphi] &\in W^{2,q_R}(D; \mathbb{R}), \quad \eta_{Niu}(\phi, u)[\hat{u}] \in W^{1,4}(D; \mathbb{R}), \\ \eta_{Di}(\phi, \partial_v u), \eta_{Di\phi'}(\phi, \partial_v u)[\varphi] &\in W^{1,q_R}(D; \mathbb{R}), \\ \eta_{Di\partial_v u}(\phi, \partial_v u)[\partial_v \hat{u}] &\in W^{2,4}(D; \mathbb{R}) \end{aligned}$$

with respect to $(\phi, u, \nabla u, \partial_v u) \in \mathcal{D} \times \mathcal{S} \times \mathcal{G} \times \mathcal{G}_{\Gamma_D}$ ($\mathcal{G} = \{\nabla u \mid u \in \mathcal{D}\}$, $\mathcal{G}_{\Gamma_D} = \{\partial_v u|_{\Gamma_D} \mid u \in \mathcal{D}\}$) and arbitrary $(\varphi, \hat{u}) \in Y \times U$. Let $\eta_{Di}(\phi, \partial_v u)$ be a linear function of $\partial_v u$. When $\eta_{Di}(\phi, \partial_v u)$ is a nonlinear function of $\partial_v u$, we assume $(\nabla \cdot \varphi)_\tau = 0$ on $\Gamma_D(\phi)$. Moreover, $(\cdot)_{\phi'}(\phi, \cdot)[\varphi]$ represents the shape derivatives of functions (Definition 9.1.1).

Moreover, if the formulae based on the partial shape derivative of a function are used, the following hypothesis will be used.

Hypothesis 9.6.2 (Cost Functions (Partial Shape Derivative)) With respect to cost function f_i ($i \in \{0, 1, \dots, m\}$) of Eq. (9.6.1), let $\zeta_i \in C^1(\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d; \mathbb{R})$, $\eta_{Ni} \in C^1(\mathbb{R}; \mathbb{R})$, $\eta_{Di} \in C^1(\mathbb{R}; \mathbb{R})$ be functions fixed in space satisfying

$$\begin{aligned} \zeta_i(\phi, u, \nabla u), \zeta_{i\phi^*}(\phi, u, \nabla u)[\varphi] &\in W^{1,q_R}(D; \mathbb{R}), \\ \zeta_{iu}(\phi, u, \nabla u)[\hat{u}] &\in L^{2q_R}(D; \mathbb{R}), \\ \zeta_{i(\nabla u)^\top}(\phi, u, \nabla u)[\nabla \hat{u}] &\in W^{1,2q_R}(D; \mathbb{R}^d), \\ \eta_{Ni}(\phi, u), \eta_{Ni\phi^*}(\phi, u)[\varphi] &\in W^{2,q_R}(D; \mathbb{R}), \quad \eta_{Niu}(\phi, u)[\hat{u}] \in W^{1,2q_R}(D; \mathbb{R}), \\ \eta_{Di}(\phi, \partial_v u), \eta_{Di\phi^*}(\phi, \partial_v u)[\varphi] &\in W^{1,q_R}(D; \mathbb{R}), \\ \eta_{Di\partial_v u}(\phi, \partial_v u)[\partial_v \hat{u}] &\in W^{2,2q_R}(D; \mathbb{R}) \end{aligned}$$

with respect to $(\phi, u, \nabla u, \partial_v u) \in \mathcal{D} \times \mathcal{S} \times \mathcal{G} \times \mathcal{G}_{\Gamma_D}$ ($\mathcal{G} = \{\nabla u \mid u \in \mathcal{D}\}$, $\mathcal{G}_{\Gamma_D} = \{\partial_v u|_{\Gamma_D} \mid u \in \mathcal{D}\}$) and arbitrary $(\varphi, \hat{u}) \in Y \times U$. Let $\eta_{Di}(\phi, \partial_v u)$ be a linear function of $\partial_v u$ and be written as $\eta_{Di\partial_v u}(\phi, \partial_v u) = v_{Di}$. Moreover, $(\cdot)_{\phi^*}(\phi, \cdot)[\varphi]$ represents the partial shape derivatives of functions (Definition 9.1.3).

These cost functions are used to define a shape optimization problem of domain variation type as follows.

Problem 9.6.3 (Shape Optimization of Domain Variation Type) Let \mathcal{D} and \mathcal{S} be defined as Eqs. (9.1.3) and (9.5.2), respectively. Also, let f_0, \dots, f_m is defined by Eq. (9.6.1). Find $\Omega(\phi)$ which satisfies

$$\min_{(\phi, u - u_D) \in \mathcal{D} \times \mathcal{S}} \left\{ f_0(\phi, u) \mid f_1(\phi, u) \leq 0, \dots, f_m(\phi, u) \leq 0, \right. \\ \left. \text{Problem 9.5.4} \right\}. \quad \square$$

In what follows, we will look at the Fréchet derivatives of cost functions and the KKT conditions with respect to a shape optimization problem (Problem 9.6.3) of domain variation type. In this respect, Lagrange functions based on several definitions will be used. Here, their relationships are summarized in order to avoid confusion. Let the Lagrange function with respect to the shape optimization problem (Problem 9.6.3) of domain variation type be

$$\mathcal{L}(\phi, u, v_0, v_1, \dots, v_m, \lambda_1, \dots, \lambda_m) \\ = \mathcal{L}_0(\phi, u, v_0) + \sum_{i \in \{1, \dots, m\}} \lambda_i \mathcal{L}_i(\phi, u, v_i), \quad (9.6.2)$$

where $\lambda = \{\lambda_1, \dots, \lambda_m\}^\top \in \mathbb{R}^m$ is a Lagrange multiplier with respect to $f_1(\phi, u) \leq 0, \dots, f_m(\phi, u) \leq 0$. Furthermore, if f_i is a functional of u for all $i \in \{0, 1, \dots, m\}$, and in view of the fact that the state determination problem (Problem 9.5.4) is an equality constraint, the functional

$$\mathcal{L}_i(\phi, u, v_i) \\ = f_i(\phi, u) + \mathcal{L}_S(\phi, u, v_i) \\ = \int_{\Omega(\phi)} (\zeta_i(\phi, u, \nabla u) - \nabla u \cdot \nabla v_i + b v_i) \, dx \\ + \int_{\Gamma_{\eta_i}(\phi)} \eta_{Ni}(\phi, u) \, d\gamma + \int_{\Gamma_p(\phi)} p_N v_i \, d\gamma \\ + \int_{\Gamma_D(\phi)} \{(u - u_D) \partial_v v_i + v_i \partial_v u - \eta_{Di}(\phi, \partial_v u)\} \, d\gamma - c_i \quad (9.6.3)$$

is called the Lagrange function of $f_i(\phi, u)$. Here, \mathcal{L}_S is the Lagrange function of the state determination problem defined by Eq. (9.5.3). Moreover, v_i is introduced as a Lagrange multiplier with respect to the state determination problem corresponding to f_i and $\tilde{v}_i = v_i - \eta_{Di} \partial_v u$ is assumed to be an element of \mathcal{S} . Similarly to u , if a variation \hat{v}_i of \tilde{v}_i is to be considered, \hat{v}_i is contained in U .

9.7 Existence of an Optimum Solution

The existence of an optimum solution of Problem 9.6.3 can be confirmed in the same fashion as in Chap. 8. To use Theorem 7.4.4 in Chap. 7, we will show the compactness of

$$\mathcal{F} = \{(\phi, u(\phi)) \in \mathcal{D} \times \mathcal{S} \mid \text{Problem 9.5.4}\} \quad (9.7.1)$$

and the continuity of f_0 . Hereinafter, we let $\tilde{u} = u - u_D \in U$.

The compactness of \mathcal{F} is presented in the following lemma [62, Lemma 2.5, p. 27, Lemma 2.15, p. 55, Lemma 2.20, p. 63].

Lemma 9.7.1 (Compactness of \mathcal{F}) *Suppose that Hypothesis 9.5.1 and Hypothesis 9.5.3 are satisfied. Moreover, $\tilde{\Gamma}_0 = \Gamma_{p0} \cup \Gamma_{\eta00} \cup \Gamma_{\eta10} \cup \dots \cup \Gamma_{\eta m0}$ is (not piecewise) $H^3 \cap C^{1,1}$ class. With respect to an arbitrary Cauchy sequence $\phi_n \rightarrow \phi$ which is uniformly convergent in \mathcal{D} and their solutions $\tilde{u}_n = \tilde{u}(\phi_n) \in U$ ($n \rightarrow \infty$) of Problem 9.5.4, the convergence*

$$\tilde{u}_n \rightarrow \tilde{u} \quad \text{strongly in } U$$

holds, and $\tilde{u} = \tilde{u}(\phi) \in U$ solves Problem 9.5.4. □

Proof Concerning the solution \tilde{u}_n of Problem 9.5.4 for ϕ_n ,

$$\alpha_n \|\tilde{u}_n\|_U^2 \leq a(\phi_n)(\tilde{u}_n, \tilde{u}_n) = \hat{l}(\phi_n)(\tilde{u}_n) \leq \|\hat{l}(\phi_n)\|_{U'} \|\tilde{u}_n\|_U$$

holds, where $a(\phi_n)$ and $\hat{l}(\phi_n)$ are defined in Eq. (9.5.4), and α_n is a positive constant used in the definition of coerciveness for $a(\phi_n)$ (see (1) in the answer to Exercise 5.2.5). When $\phi_n \rightarrow \phi$ is uniformly convergent in \mathcal{D} , α_n can be replaced by a positive constant α not depending on n . The norm $\|\hat{l}(\phi_n)\|_{U'} = \|l(\phi_n) + a(\phi_n)(u_D, \cdot)\|_{U'}$ ($l(\phi_n)$ defined in Eq. (9.5.4)) being bounded can be shown using (3) in the answer to Exercise 5.2.5 by replacing $\hat{l}(v)$ and Ω in Exercise 5.2.5 by $\hat{l}(\phi_n)(v)$ and $\Omega(\phi_n)$, respectively. Hence, there exists a subsequence such that $\tilde{u}_n \rightarrow \tilde{u}$ weakly in U .

Next, we will show that \tilde{u} solves Problem 9.5.4 for ϕ . From the definition of Problem 9.5.4,

$$\lim_{n \rightarrow \infty} a(\phi_n)(\tilde{u}_n, v) = \lim_{n \rightarrow \infty} \hat{l}(\phi_n)(v) \quad (9.7.2)$$

holds with respect to an arbitrary $v \in U$. From Hypothesis 9.5.2, the right-hand side of Eq. (9.7.2) becomes

$$\lim_{n \rightarrow \infty} \hat{l}(\phi_n)(v) = \hat{l}(\phi)(v). \quad (9.7.3)$$

Indeed,

$$\begin{aligned}
& \left| \hat{l}(\boldsymbol{\phi}_n)(v) - \hat{l}(\boldsymbol{\phi})(v) \right| \\
& \leq \left| \int_{\Omega(\boldsymbol{\phi}_n)} b(\boldsymbol{\phi}_n) v \, dx - \int_{\Omega(\boldsymbol{\phi})} b(\boldsymbol{\phi}) v \, dx \right| \\
& \quad + \left| \int_{\Gamma_p(\boldsymbol{\phi}_n)} p_N(\boldsymbol{\phi}_n) v \, d\gamma - \int_{\Gamma_p(\boldsymbol{\phi})} p_N(\boldsymbol{\phi}) v \, d\gamma \right| \\
& \quad + \left| \int_{\Omega(\boldsymbol{\phi}_n)} \nabla u_D(\boldsymbol{\phi}_n) \cdot \nabla v \, dx - \int_{\Omega(\boldsymbol{\phi})} \nabla u_D(\boldsymbol{\phi}) \cdot \nabla v \, dx \right|
\end{aligned} \tag{9.7.4}$$

holds. The first term in the right-hand side of Eq. (9.7.4) becomes

$$\begin{aligned}
& \left| \int_D \left(\chi_{\Omega(\boldsymbol{\phi}_n)} b(\boldsymbol{\phi}_n) - \chi_{\Omega(\boldsymbol{\phi})} b(\boldsymbol{\phi}) \right) v \, dx \right| \\
& \leq \left| \int_D \chi_{\Omega(\boldsymbol{\phi})} (b(\boldsymbol{\phi}_n) - b(\boldsymbol{\phi})) v \, dx \right| + \left| \int_D (\chi_{\Omega(\boldsymbol{\phi}_n)} - \chi_{\Omega(\boldsymbol{\phi})}) b(\boldsymbol{\phi}_n) v \, dx \right|,
\end{aligned}$$

where χ_{Ω} denotes the characteristic function such that $\chi_{\Omega} : D \rightarrow \mathbb{R}$ ($\chi_{\Omega}(\Omega) = 1$, $\chi_{\Omega}(D \setminus \bar{\Omega}) = 0$). Using $b \in C_{S'}^1(B; C^{0,1}(D; \mathbb{R}))$ in Hypothesis 9.5.1 and the property [67, Proposition 2.2.28, p. 45]

$$\chi_{\Omega(\boldsymbol{\phi}_n)} \rightarrow \chi_{\Omega(\boldsymbol{\phi})} \quad \text{in } L^\infty(D; \mathbb{R})\text{-weak}^*, \tag{9.7.5}$$

the first term in the right-hand side of Eq. (9.7.4) converges to zero. It can also be shown that the third term in the right-hand side of Eq. (9.7.4) converges to zero using $u_D \in C_{S'}^1(B; W^{2,4}(D; \mathbb{R}))$ and Eq. (9.7.5).

The convergence to zero of the second term in the right-hand side of Eq. (9.7.4) can be confirmed in the following way. Here, we modify the condition for $\Gamma_p(\boldsymbol{\phi}_n)$ in \mathcal{D} defined in Eq. (9.1.3) (a class of $H^3 \cap C^{1,1}$) as follows. $\Gamma_p(\boldsymbol{\phi}_n)$ can be defined using a function $\boldsymbol{\sigma}(\boldsymbol{\phi}_n)(\xi) = \boldsymbol{\sigma}_n(\xi)$ of a parameter $\xi \in \Xi = (0, 1)^{d-1}$ as

$$\begin{aligned}
\Gamma_p(\boldsymbol{\phi}_n) &= \tilde{\Gamma}_p(\boldsymbol{\sigma}_n) \\
&= \left\{ \boldsymbol{\sigma}_n \in H^3 \cap C^{1,1}(\Xi; \mathbb{R}^d) \mid \|\boldsymbol{\sigma}_n\|_{\mathbb{R}^d} \leq c_0, \right. \\
&\quad c_1 \leq \left\| \nabla_{\xi} \boldsymbol{\sigma}_n^{\top} \right\|_{\mathbb{R}^{(d-1) \times d}} \leq c_2, \\
&\quad \left. \left\| \nabla_{\xi}^{|\boldsymbol{\beta}|} \boldsymbol{\sigma}_n^{\top} \right\|_{\mathbb{R}^{(d-1)^2 \times d}} \leq c_3 \quad (|\boldsymbol{\beta}| = 2) \text{ a.e. in } \Xi \right\},
\end{aligned} \tag{9.7.6}$$

where $\nabla_\xi = (\partial/\partial\xi_i)_i$ and c_0, \dots, c_3 are positive constants. Hereafter, $\omega_{\Xi n}$ and ω_Ξ denote $\|\nabla_\xi \sigma_n^\top\|_{\mathbb{R}^{(d-1)\times d}}$ and $\|\nabla_\xi \sigma^\top\|_{\mathbb{R}^{(d-1)\times d}}$, respectively. Moreover, let $\tilde{p}_N(t) = p_N(t\phi_n + (1-t)\phi)$ ($t \in [0, 1]$). Here, using $\phi_n \rightarrow \phi$ (uniformly convergent in \mathcal{D}), boundedness of the trace operator $\|\gamma_{\Gamma_p(\phi)}\|$ (Eq. (5.2.4)), the result in [29, Corollary 1] and $p_N \in C_{S'}^1(B; C^{1,1}(D; \mathbb{R}))$ in Hypothesis 9.5.1, the second term of the right-hand side of Eq. (9.7.4) becomes

$$\begin{aligned}
& \left| \int_{\tilde{\Gamma}_p(\sigma_n)} p_N(\phi_n) v \, d\gamma - \int_{\tilde{\Gamma}_p(\sigma)} p_N(\phi) v \, d\gamma \right| \\
& \leq \left| \int_{\Xi} \{ (p_N(\phi_n) \circ \sigma_n) (v \circ \sigma_n) \omega_{\Xi n} - (p_N(\phi) \circ \sigma) (v \circ \sigma) \omega_\Xi \} \, d\sigma \right| \\
& \leq \left| \int_{\Xi} \{ (p_N(\phi_n) \circ \sigma_n) - (p_N(\phi_n) \circ \sigma) \} (v \circ \sigma_n) \omega_{\Xi n} \, d\sigma \right| \\
& \quad + \left| \int_{\Xi} \{ (p_N(\phi_n) \circ \sigma) - (p_N(\phi) \circ \sigma) \} (v \circ \sigma_n) \omega_{\Xi n} \, d\sigma \right| \\
& \quad + \left| \int_{\Xi} (p_N(\phi) \circ \sigma) (v \circ \sigma_n - v \circ \sigma) \omega_{\Xi n} \, d\sigma \right| \\
& \quad + \left| \int_{\Xi} (p_N(\phi) \circ \sigma) (v \circ \sigma) (\omega_{\Xi n} - \omega_\Xi) \, d\sigma \right| \\
& \leq \sqrt{c_2} \|v\|_{L^2(\Gamma_p(\phi_n); \mathbb{R})} \| (p_N(\phi_n) \circ \sigma_n) - (p_N(\phi_n) \circ \sigma) \|_{L^2(\Xi; \mathbb{R})} \\
& \quad + \sqrt{\frac{c_2}{c_1}} \|v\|_{L^2(\Gamma_p(\phi_n); \mathbb{R})} \|p_N(\phi_n) - p_N(\phi)\|_{L^2(\Gamma_p(\phi); \mathbb{R})} \\
& \quad + \sqrt{c_2} \|p_N(\phi)\|_{L^2(\Gamma_p(\phi); \mathbb{R})} \|v \circ \sigma_n - v \circ \sigma\|_{L^2(\Xi; \mathbb{R})} \\
& \quad + \frac{1}{c_1} \|\omega_{\Xi n} - \omega_\Xi\|_{H^2 \cap C^{0,1}(\Xi; \mathbb{R}^d)} \|p_N(\phi)\|_{L^2(\Gamma_p(\phi); \mathbb{R})} \|v\|_{L^2(\Gamma_p(\phi); \mathbb{R})} \\
& \leq \sqrt{c_2} \|\gamma_{\Gamma_p(\phi)}\|^2 \|v\|_U \|p_N(\phi_n)\|_{C^{1,1}(D; \mathbb{R})} \|\sigma_n - \sigma\|_{H^3 \cap C^{1,1}(\Xi; \mathbb{R}^d)} \\
& \quad + \sqrt{\frac{c_2}{c_1}} \|\gamma_{\Gamma_p(\phi)}\|^2 \|v\|_U \sup_{t \in [0, 1]} \|\tilde{p}_N'(t)\|_{C^{1,1}(D; \mathbb{R})} \|\phi_n - \phi\|_X \\
& \quad + \sqrt{c_2} \|\gamma_{\Gamma_p(\phi)}\|^2 \|p_N(\phi)\|_{C^{1,1}(D; \mathbb{R})} \|v\|_U \|\sigma_n - \sigma\|_{H^3 \cap C^{1,1}(\Xi; \mathbb{R}^d)} \\
& \quad + \frac{1}{c_1} \|\gamma_{\Gamma_p(\phi)}\|^2 \|\omega_{\Xi n} - \omega_\Xi\|_{H^2 \cap C^{0,1}(\Xi; \mathbb{R}^d)} \|p_N(\phi)\|_{C^{1,1}(D; \mathbb{R})} \|v\|_U \\
& \rightarrow 0 \quad (n \rightarrow \infty). \tag{9.7.7}
\end{aligned}$$

In Eq. (9.7.7), we used the relations

$$\begin{aligned} \left| \int_{\Xi} (v \circ \sigma) \omega_{\Xi} d\sigma \right| &\leq \sqrt{c_2} \left(\int_{\Xi} (v \circ \sigma)^2 \omega_{\Xi} d\sigma \right)^{1/2} = \sqrt{c_2} \|v\|_{L^2(\Gamma_p(\phi_n); \mathbb{R})}, \\ \left| \int_{\Xi} \{(p_N(\phi_n) \circ \sigma) - (p_N(\phi) \circ \sigma)\} d\sigma \right| \\ &\leq \frac{1}{\sqrt{c_1}} \left(\int_{\Xi} \{(p_N(\phi_n) \circ \sigma) - (p_N(\phi) \circ \sigma)\}^2 \omega_{\Xi} d\sigma \right)^{1/2} \\ &= \frac{1}{\sqrt{c_1}} \|p_N(\phi_n) - p_N(\phi)\|_{L^2(\Gamma_p(\phi); \mathbb{R})}. \end{aligned}$$

Using the results above, Eq. (9.7.3) is proved.

The left-hand side of Eq. (9.7.2) becomes

$$\lim_{n \rightarrow \infty} a(\phi_n)(u_n, v) = a(\phi)(u, v). \quad (9.7.8)$$

It can be confirmed by

$$\begin{aligned} &|a(\phi_n)(\tilde{u}_n, v) - a(\phi)(\tilde{u}, v)| \\ &= \left| \int_{\Omega(\phi_n)} \nabla \tilde{u}_n \cdot \nabla v \, dx - \int_{\Omega(\phi)} \nabla \tilde{u} \cdot \nabla v \, dx \right| \\ &= \left| \int_D (\chi_{\Omega(\phi_n)} \nabla \tilde{u}_n - \chi_{\Omega(\phi)} \nabla \tilde{u}) \cdot \nabla v \, dx \right| \\ &\leq \left| \int_D \chi_{\Omega(\phi)} (\nabla \tilde{u}_n - \nabla \tilde{u}) \cdot \nabla v \, dx \right| \\ &\quad + \left| \int_D (\chi_{\Omega(\phi_n)} - \chi_{\Omega(\phi)}) \nabla \tilde{u}_n \cdot \nabla v \, dx \right|. \end{aligned} \quad (9.7.9)$$

To the right-hand side of Eq. (9.7.9), we adopt $\tilde{u}_n \rightarrow \tilde{u}$ weakly in U and the property Eq. (9.7.5) for the characteristic function, and obtain Eq. (9.7.8). Substituting Eqs. (9.7.3) and (9.7.8) into Eq. (9.7.2), the weak form of Problem 9.5.4 can be obtained. Namely, $\tilde{u} = \tilde{u}(\phi) \in U$ is the solution of Problem 9.5.4.

Since the weak convergence was shown, then to prove the strong convergence of $\{u_n\}_{n \in \mathbb{N}}$ to u , it is sufficient to show that

$$\|u_n\|_U \rightarrow \|u\|_U \quad (n \rightarrow \infty). \quad (9.7.10)$$

Indeed, when using $a(\phi)$ in Eq. (9.5.5) and taking

$$\|v\| = a(\phi)(v, v)$$

as a norm on U , we have

$$\begin{aligned} \|u_n\| &= a(\phi)(u_n, u_n) = \int_D \left(\chi_{\Omega(\phi)} - \chi_{\Omega(\phi_n)} \right) \nabla u_n \cdot \nabla u_n \, dx + a(\phi_n)(u_n, u_n) \\ &= \int_D \left(\chi_{\Omega(\phi)} - \chi_{\Omega(\phi_n)} \right) \nabla u_n \cdot \nabla u_n \, dx + l(\phi_n)(u_n) \\ &\rightarrow l(\phi)(u) = \|u\| \quad (n \rightarrow \infty). \end{aligned} \quad (9.7.11)$$

Then, $u_n \rightarrow u$ strongly in U is proved. \square

We consider that the condition of $\tilde{u}(\phi)$ included in \mathcal{S} is guaranteed in the setting of Problem 9.5.4 satisfying Hypotheses 9.5.1 and 9.5.3.

The latter assumption in Theorem 7.4.4 (continuity of f_0) means that f_0 is continuous on

$$S = \{(\phi, \tilde{u}(\phi)) \in \mathcal{F} \mid f_1(\phi, u(\phi)) \leq 0, \dots, f_m(\phi, u(\phi)) \leq 0\}. \quad (9.7.12)$$

S depends on the problem setting. Then, we will confirm the continuity of f_0 by showing the continuity of f_i ($i \in \{0, 1, \dots, m\}$) by the following lemma and assuming that S is not empty.

Lemma 9.7.2 (Continuity of f_i) *Let f_i be defined as in Eq. (9.6.1) under Hypothesis 9.6.1. Let $u_n \rightarrow u$ strongly in U be determined by Lemma 9.7.1 with respect to an arbitrary Cauchy sequence $\phi_n \rightarrow \phi$ in X which is uniformly convergent in \mathcal{D} , and satisfy $\|\partial_v u_n - \partial_v u\|_{L^2(\Gamma_D; \mathbb{R})} \rightarrow 0$ ($n \rightarrow \infty$) on Γ_D . Then, f_i is continuous with respect to $\phi \in \mathcal{D}$.* \square

Proof The proof will be completed when

$$\begin{aligned} &|f_i(\phi_n, u_n) - f_i(\phi, u)| \\ &\leq \left| \int_{\Omega(\phi_n)} \zeta_i(\phi_n, u_n, \nabla u_n) \, dx - \int_{\Omega(\phi)} \zeta_i(\phi, u, \nabla u) \, dx \right| \\ &\quad + \left| \int_{\Gamma_{\eta i}(\phi_n)} \eta_{Ni}(\phi_n, u_n) \, d\gamma - \int_{\Gamma_{\eta i}(\phi)} \eta_{Ni}(\phi, u) \, d\gamma \right| \\ &\quad + \left| \int_{\Gamma_D(\phi_n)} \eta_{Di}(\phi_n, \partial_v u_n) \, d\gamma - \int_{\Gamma_D(\phi)} \eta_{Di}(\phi, \partial_v u) \, d\gamma \right| \\ &= e_{\Omega} + e_{\Gamma_{\eta}} + e_{\Gamma_D} \rightarrow 0 \quad (n \rightarrow \infty) \end{aligned} \quad (9.7.13)$$

is shown with respect to $\phi_n \rightarrow \phi$ which is uniformly convergent in \mathcal{D} . For e_Ω ,

$$\begin{aligned} e_\Omega &\leq \left| \int_D \left(\chi_{\Omega(\phi_n)} - \chi_{\Omega(\phi)} \right) \zeta_i(\phi_n, u_n, \nabla u_n) dx \right| \\ &\quad + \left| \int_D \chi_{\Omega(\phi)} (\zeta_i(\phi_n, u_n, \nabla u_n) - \zeta_i(\phi, u, \nabla u)) dx \right| \\ &= e_{\Omega 1} + e_{\Omega 2} \end{aligned}$$

holds. $e_{\Omega 1}$ converges to zero by Eq. (9.7.5). For $e_{\Omega 2}$, using $\tilde{u}_n \rightarrow \tilde{u}$ weakly in U and ζ_i , notation $\tilde{\zeta}_i(t) = \zeta_i(t\phi_n + (1-t)\phi, tu_n + (1-t)n, t\nabla u_n + (1-t)\nabla u)$ ($t \in [0, 1]$) and $\zeta_i \in C^1(\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d; \mathbb{R})$ in Hypothesis 9.6.1,

$$\begin{aligned} e_{\Omega 2} &\leq \sup_{t \in [0, 1]} \left| \int_{\Omega(\phi)} \tilde{\zeta}_{i\phi'}(t) [\phi_n - \phi] dx \right| + \sup_{t \in [0, 1]} \left| \int_{\Omega(\phi)} \tilde{\zeta}_{iu}(t) [u_n - u] dx \right| \\ &\quad + \sup_{t \in [0, 1]} \left| \int_{\Omega(\phi)} \tilde{\zeta}_i \nabla u(t) [\nabla u_n - \nabla u] dx \right| \\ &\leq \sup_{t \in [0, 1]} \left\| \tilde{\zeta}_{i\phi'}(t) \right\|_{H^1 \cap L^\infty(D; \mathbb{R})} \|\phi_n - \phi\|_X \\ &\quad + \sup_{t \in [0, 1]} \left\| \tilde{\zeta}_{iu}(t) \right\|_{L^4(D; \mathbb{R})} \|u_n - u\|_U \\ &\quad + \sup_{t \in [0, 1]} \left\| \tilde{\zeta}_i \nabla u(t) \right\|_{W^{1,4}(D; \mathbb{R})} \|\nabla u_n - \nabla u\|_{L^2(D; \mathbb{R})} \\ &\rightarrow 0 \quad (n \rightarrow \infty) \end{aligned}$$

holds. The convergence of e_{Γ_η} to zero can be shown as follows. Assuming a similar condition to Eq. (9.7.6) for $\Gamma_{\eta i}(\phi_n)$, $\Gamma_{\eta i}(\phi_n)$ can be represented with the parameter $\sigma_n(\xi)$ ($\xi \in \Xi = (0, 1)^{d-1}$) as $\tilde{\Gamma}_{\eta i}(\sigma_n)$. Using notation $\tilde{\eta}_{Ni}(t) = \eta_{Ni}(t\phi_n + (1-t)\phi, tu_n + (1-t)u)$ ($t \in [0, 1]$), $\tilde{u}_n \rightarrow \tilde{u}$ weakly in U , boundedness of the trace operator, the result in [29, Corollary 1] and $\eta_{Ni} \in W^{2,q_R}(D; \mathbb{R})$ in Hypothesis 9.6.1, we have

$$\begin{aligned} e_{\Gamma_\eta} &= \left| \int_{\tilde{\Gamma}_{\eta i}(\sigma_n)} \eta_{Ni}(\phi_n, u_n) d\gamma - \int_{\tilde{\Gamma}_{\eta i}(\sigma)} \eta_{Ni}(\phi, u) d\gamma \right| \\ &\leq \left| \int_{\Xi} \{ (\eta_{Ni}(\phi_n, u_n) \circ \sigma_n) \omega_{\Xi n} - (\eta_{Ni}(\phi, u) \circ \sigma) \omega_{\Xi} \} d\sigma \right| \\ &\leq \left| \int_{\Xi} \{ (\eta_{Ni}(\phi_n, u_n) \circ \sigma_n) - (\eta_{Ni}(\phi_n, u_n) \circ \sigma) \} \omega_{\Xi n} d\sigma \right| \end{aligned}$$

$$\begin{aligned}
& + \left| \int_{\Xi} \{ (\eta_{Ni}(\phi_n, u_n) \circ \sigma) - (\eta_{Ni}(\phi, u_n) \circ \sigma) \} \omega_{\Xi n} \, d\sigma \right| \\
& + \left| \int_{\Xi} \{ (\eta_{Ni}(\phi, u_n) \circ \sigma) - (\eta_{Ni}(\phi, u) \circ \sigma) \} \omega_{\Xi n} \, d\sigma \right| \\
& + \left| \int_{\Xi} (\eta_{Ni}(\phi, u) \circ \sigma) (\omega_{\Xi n} - \omega_{\Xi}) \, d\sigma \right| \\
& \leq \sqrt{c_2} \| (\eta_{Ni}(\phi_n, u_n) \circ \sigma_n) - (\eta_{Ni}(\phi_n, u_n) \circ \sigma) \|_{L^2(\Xi; \mathbb{R})} \\
& \quad + \sqrt{\frac{c_2}{c_1}} \| \eta_{Ni}(\phi_n, u_n) - \eta_{Ni}(\phi, u) \|_{L^2(\Gamma_{\eta i}(\phi); \mathbb{R})} \\
& \quad + \sqrt{\frac{c_2}{c_1}} \| \eta_{Ni}(\phi_n, u_n) - \eta_{Ni}(\phi, u) \|_{L^2(\Gamma_{\eta i}(\phi); \mathbb{R})} \\
& \quad + \frac{1}{c_1} \| \omega_{\Xi n} - \omega_{\Xi} \|_{H^2 \cap C^{0,1}(\Xi; \mathbb{R}^d)} \| \eta_{Ni}(\phi, u) \|_{L^2(\Gamma_{\eta i}(\phi); \mathbb{R})} \\
& \leq \sqrt{c_2} \| \gamma_{\Gamma_{\eta i}(\phi)} \| \| \eta_{Ni}(\phi_n, u_n) \|_{W^{2,q_R}(D; \mathbb{R})} \| \sigma_n - \sigma \|_{C^{1,1}(\Xi; \mathbb{R}^d)} \\
& \quad + \sqrt{\frac{c_2}{c_1}} \| \gamma_{\Gamma_{\eta i}(\phi)} \| \sup_{t \in [0, 1]} \| \tilde{\eta}_{Ni} \phi(t) \|_{W^{2,q_R}(D; \mathbb{R})} \| \phi_n - \phi \|_X \\
& \quad + \sqrt{\frac{c_2}{c_1}} \| \gamma_{\Gamma_{\eta i}(\phi)} \| \sup_{t \in [0, 1]} \| \tilde{\eta}_{Ni} u(t) \|_{W^{2,q_R}(D; \mathbb{R})} \| u_n - u \|_U \\
& \quad + \frac{1}{c_1} \| \gamma_{\Gamma_{\eta i}(\phi)} \| \| \omega_{\Xi n} - \omega_{\Xi} \|_{H^2 \cap C^{0,1}(\Xi; \mathbb{R}^d)} \| \eta_{Ni}(\phi, u) \|_{W^{2,q_R}(D; \mathbb{R})} \\
& \rightarrow 0 \quad (n \rightarrow \infty), \tag{9.7.14}
\end{aligned}$$

where c_1 and c_2 are positive constants when Eq. (9.7.6) is rewritten for $\Gamma_{\eta i}(\phi_n)$. $e_{\Gamma_D} \rightarrow 0$ ($n \rightarrow \infty$) can be shown using $\| \partial_v u_n - \partial_v u \|_{L^2(\Gamma_D; \mathbb{R})} \rightarrow 0$ ($n \rightarrow \infty$) in a similar way. Based on the results above, Eq. (9.7.13) is shown. \square

In Theorem 7.4.4 showing the existence of a solution in the abstract optimum design problem, the first assumption (compactness of \mathcal{F}) was confirmed by Lemma 9.7.1. The second assumption (continuity of f_0) can be satisfied with the conditions for Lemma 9.7.2 and the assumption that S is not empty. Then, under the conditions, it can be assured that there exists an optimum solution of Problem 9.6.3.

Regarding the solution of Problem 9.6.3, let us recall the similar situation of Remark 8.4.3 in Chap. 8. In the definition of \mathcal{D} shown in Eq. (9.1.3), a side constraint $\| \phi \|_{H^2 \cap C^{0,1}(D; \mathbb{R}^d)} \leq \beta$ is added. When this condition becomes active, we have to deal this condition as an inequality condition. Depending on the setting of the problem, we may meet a situation such that a boundary converges to a shape with sharp corners which is not a Lipschitz boundary. In this case, a converged shape can

be obtained by activating the side constraint. Moreover, regarding the selection of X and \mathcal{D} , the same situation as Remark 8.4.4 holds.

9.8 Derivatives of Cost Functions

In this chapter, we consider the solution of the shape optimization problem (Problem 9.6.3) of domain variation type using a gradient method and a Newton method. In order to use the gradient method, the first-order shape derivatives of cost functions are necessary. Moreover, if the Newton method is to be used, the second-order shape derivatives (Hessians) of the cost functions are required. Here, let us obtain the first and second-order shape derivatives of the cost functions f_i using the Lagrange multiplier method shown in Sect. 7.5.2 and the method shown in 7.5.3, respectively. In this case, let us look at the methods using the formulae based on the shape derivative of a function separately from the method using the formulae based on the partial shape derivative of a function shown in Sect. 9.3. However, with respect to the second-order shape derivatives, only the results using the method with the formulae based on the shape derivative of a function will be shown.

9.8.1 Shape Derivative of f_i Using Formulae Based on Shape Derivative of a Function

Firstly, let us use the formulae based on the shape derivative of a function (Sect. 9.3.1) to obtain the Fréchet derivative of \mathcal{L}_i and use its stationary conditions to seek the shape derivative of f_i .

The Fréchet derivative of $\mathcal{L}_i(\boldsymbol{\phi}, u, v_i)$ is written as

$$\begin{aligned} \mathcal{L}'_i(\boldsymbol{\phi}, u, v_i) & [\boldsymbol{\varphi}, \hat{u}, \hat{v}_i] \\ &= \mathcal{L}_{i\boldsymbol{\varphi}}(\boldsymbol{\phi}, u, v_i) [\boldsymbol{\varphi}] + \mathcal{L}_{iu}(\boldsymbol{\phi}, u, v_i) [\hat{u}] + \mathcal{L}_{iv_i}(\boldsymbol{\phi}, u, v_i) [\hat{v}_i] \end{aligned} \quad (9.8.1)$$

with respect to an arbitrary $(\boldsymbol{\varphi}, \hat{u}, \hat{v}_i) \in X \times U \times U$. Here, the notations in Eqs. (9.3.5) and (9.3.15) are used. In this case, the shape derivative u' used in Eqs. (9.3.5) and (9.3.15) following Definition 9.1.1 was replaced with an arbitrary $\hat{u} \in X$, because it was assumed that u is not necessarily the solution of Problem 9.5.4 in the definition of the Lagrange function. Let us look at each term in detail below.

The third term on the right-hand side of Eq. (9.8.1) becomes

$$\mathcal{L}_{iv_i}(\boldsymbol{\phi}, u, v_i) [\hat{v}_i] = \mathcal{L}_{Sv_i}(\boldsymbol{\phi}, u, v_i) [\hat{v}_i] = \mathcal{L}_S(\boldsymbol{\phi}, u, \hat{v}_i). \quad (9.8.2)$$

Equation (9.8.2) is the Lagrange function of the state determination problem (Problem 9.5.4). Hence, if u is a weak solution of the state determination problem, its term is zero.

Moreover, the second term on the right-hand side of Eq. (9.8.1) becomes

$$\begin{aligned}
 \mathcal{L}_{iu}(\boldsymbol{\phi}, u, v_i)[\hat{u}] &= \int_{\Omega(\boldsymbol{\phi})} (-\nabla \hat{u} \cdot \nabla v_i + \zeta_{iu}(\boldsymbol{\phi}, u, \nabla u) \hat{u} + \zeta_{i(\nabla u)^\top}(\boldsymbol{\phi}, u, \nabla u) \nabla \hat{u}) dx \\
 &\quad + \int_{\Gamma_{\eta i}(\boldsymbol{\phi})} \eta_{Ni u}(\boldsymbol{\phi}, u) \hat{u} d\gamma \\
 &\quad + \int_{\Gamma_D(\boldsymbol{\phi})} \{\hat{u} \partial_v v_i + (v_i - \eta_{Di \partial_v u}(\boldsymbol{\phi}, \partial_v u)) \partial_v \hat{u}\} d\gamma. \tag{9.8.3}
 \end{aligned}$$

Here, if v_i can be determined so that Eq. (9.8.3) equates to zero with respect to an arbitrary $\hat{u} \in U$, the second term on the right-hand side of Eq. (9.8.1) also vanishes. From the fact that

$$\begin{aligned}
 &\int_{\Omega(\boldsymbol{\phi})} \left(\zeta_{i(\nabla u)^\top}(u, \nabla u) \nabla \hat{u} - \nabla \hat{u} \cdot \nabla v_i \right) dx \\
 &= \int_{\partial \Omega(\boldsymbol{\phi})} \hat{u} \left(\zeta_{i(\nabla u)^\top} - \nabla v_i \right) \cdot \mathbf{v} d\gamma - \int_{\Omega(\boldsymbol{\phi})} \hat{u} \nabla \cdot \left(\zeta_{i(\nabla u)^\top} - \nabla v_i \right) dx
 \end{aligned}$$

holds if $v_i \in W^{2,4}(D; \mathbb{R})$ is assumed, its strong form can be written as follows.

Problem 9.8.1 (Adjoint Problem with Respect to f_i) When the solution u to Problem 9.5.4 with respect to $\boldsymbol{\phi} \in \mathcal{D}$ is obtained, find $v_i : \Omega(\boldsymbol{\phi}) \rightarrow \mathbb{R}$ which satisfies

$$\begin{aligned}
 -\Delta v_i &= \zeta_{iu}(\boldsymbol{\phi}, u, \nabla u) - \nabla \cdot \zeta_{i(\nabla u)^\top}(\boldsymbol{\phi}, u, \nabla u) \quad \text{in } \Omega(\boldsymbol{\phi}), \\
 \partial_v v_i &= \eta_{Ni u}(\boldsymbol{\phi}, u) + \zeta_{i(\nabla u)^\top}(\boldsymbol{\phi}, u, \nabla u) \cdot \mathbf{v} \quad \text{on } \Gamma_{\eta i}(\boldsymbol{\phi}), \\
 \partial_v v_i &= \zeta_{i(\nabla u)^\top}(\boldsymbol{\phi}, u, \nabla u) \cdot \mathbf{v} \quad \text{on } \Gamma_N(\boldsymbol{\phi}) \setminus \bar{\Gamma}_{\eta i}(\boldsymbol{\phi}), \\
 v_i &= \eta_{Di \partial_v u}(\boldsymbol{\phi}, \partial_v u) \quad \text{on } \Gamma_D(\boldsymbol{\phi}). \tag*{\square}
 \end{aligned}$$

Here, the admissible set of adjoint variables for $v_i - \eta_{Di \partial_v u}$ is taken to be \mathcal{S} in order to obtain a regular solution of the shape optimization problem of domain variation type. In Hypothesis 9.6.1, the regularities for ζ_{iu} , $\zeta_{i(\nabla u)^\top}$, $\eta_{Ni u}$ and $\eta_{Di \partial_v u}$ were given to obtain this result.

Furthermore, the first term on the right-hand side of Eq. (9.8.1) becomes

$$\begin{aligned}
& \mathcal{L}_{i\phi'}(\phi, u, v_i)[\varphi] \\
&= \int_{\Omega(\phi)} \left[\nabla u \cdot \left\{ (\nabla \varphi^\top) \nabla v_i \right\} + \nabla v_i \cdot \left\{ (\nabla \varphi^\top) \nabla u \right\} \right. \\
&\quad \left. - \xi_i (\nabla u)^\top \cdot \left\{ (\nabla \varphi^\top) \nabla u \right\} + (\xi_i - \nabla u \cdot \nabla v_i + b v_i) \nabla \cdot \varphi \right. \\
&\quad \left. + (\xi_{i\phi'} + u b') \cdot \varphi \right] dx \\
&+ \int_{\Gamma_{\eta i}(\phi)} (\kappa \eta_{Ni} \mathbf{v} \cdot \varphi - \nabla_\tau \eta_{Ni} \cdot \varphi_\tau + \eta_{Ni\phi'} \cdot \varphi) d\gamma \\
&+ \int_{\partial \Gamma_{\eta i}(\phi) \cup \Theta_{\eta i}(\phi)} \eta_{Ni} \boldsymbol{\tau} \cdot \varphi d\zeta \\
&+ \int_{\Gamma_p(\phi)} \left\{ \kappa p_N v_i \mathbf{v} \cdot \varphi - \nabla_\tau (p_N v_i) \cdot \varphi_\tau + v_i p'_N \cdot \varphi \right\} d\gamma \\
&+ \int_{\partial \Gamma_p(\phi) \cup \Theta_p(\phi)} p_N v_i \boldsymbol{\tau} \cdot \varphi d\zeta \\
&+ \int_{\Gamma_D(\phi)} \left[\left\{ (u - u_D) w(\varphi, v_i) + (v_i - \eta_{Di} \partial_v u) w(\varphi, u) \right\} \right. \\
&\quad \left. + \{(u - u_D) \partial_v v_i + v_i \partial_v u - \eta_{Di}(\phi, \partial_v u)\} (\nabla \cdot \varphi)_\tau + \eta_{Di\phi'} \cdot \varphi \right] d\gamma
\end{aligned} \tag{9.8.4}$$

using the formulae of Eq. (9.3.5), representing the result of Proposition 9.3.4, and Eq. (9.3.15), representing the result of Proposition 9.3.7. Here, $w(\varphi, u)$ and $(\nabla \cdot \varphi)_\tau$ follow Eqs. (9.3.12) and (9.2.6), respectively. Moreover, the fact that $\Gamma_p(\phi)$ and $\Gamma_{\eta i}(\phi)$ are piecewise $H^3 \cap C^{1,1}$ (assumed in the definition of \mathcal{D}) was used to obtain the integral on $\Gamma_p(\phi)$ and $\Gamma_{\eta i}(\phi)$.

Bearing the above results in mind, when u and v_i are the weak solutions of Problem 9.5.4 and Problem 9.8.1, respectively, and the Dirichlet conditions corresponding to these problems, as well as the condition for η_{Di} in Hypothesis 9.6.1 hold, the integral on $\Gamma_D(\phi)$ on Eq. (9.8.4) will be zero except the term of $\eta_{Di\phi'} \cdot \varphi$. Hence, using the notation of Eq. (7.5.15) for \tilde{f}_i , we obtain

$$\begin{aligned}
\tilde{f}'_i(\phi)[\varphi] &= \mathcal{L}_{i\phi'}(\phi, u, v_i)[\varphi] = \langle \mathbf{g}_i, \varphi \rangle \\
&= \int_{\Omega(\phi)} \left\{ \mathbf{G}_{\Omega i} \cdot (\nabla \varphi^\top) + g_{\Omega i} \nabla \cdot \varphi + \mathbf{g}_{\zeta bi} \cdot \varphi \right\} dx \\
&+ \int_{\Gamma_p(\phi)} \mathbf{g}_{pi} \cdot \varphi d\gamma + \int_{\partial \Gamma_p(\phi) \cup \Theta_p(\phi)} \mathbf{g}_{\partial pi} \cdot \varphi d\zeta
\end{aligned}$$

$$\begin{aligned}
& + \int_{\Gamma_{\eta i}(\phi)} \mathbf{g}_{\eta i} \cdot \boldsymbol{\varphi} \, d\gamma + \int_{\partial\Gamma_{\eta i}(\phi) \cup \Theta_{\eta i}(\phi)} \mathbf{g}_{\partial\eta i} \cdot \boldsymbol{\varphi} \, d\varsigma \\
& + \int_{\Gamma_D(\phi)} \mathbf{g}_{D i} \cdot \boldsymbol{\varphi} \, d\gamma,
\end{aligned} \tag{9.8.5}$$

where

$$\mathbf{G}_{\Omega i} = \nabla u (\nabla v_i)^\top + \nabla v_i (\nabla u)^\top - \zeta_i (\nabla u)^\top (\nabla u)^\top, \tag{9.8.6}$$

$$g_{\Omega i} = \zeta_i - \nabla u \cdot \nabla v_i + b v_i, \tag{9.8.7}$$

$$\mathbf{g}_{\zeta b i} = \zeta_i \phi' + u b', \tag{9.8.8}$$

$$\mathbf{g}_{p i} = \kappa p_N v_i \mathbf{v} - \sum_{j \in \{1, \dots, d-1\}} \{ \boldsymbol{\tau}_j \cdot \nabla (p_N v_i) \} \boldsymbol{\tau}_j + v_i p'_N, \tag{9.8.9}$$

$$\mathbf{g}_{\partial p i} = p_N v_i \boldsymbol{\tau}, \tag{9.8.10}$$

$$\mathbf{g}_{\eta i} = \kappa \eta_{N i} \mathbf{v} - \sum_{j \in \{1, \dots, d-1\}} (\boldsymbol{\tau}_j \cdot \nabla \eta_{N i}) \boldsymbol{\tau}_j + \eta_{N i} \phi', \tag{9.8.11}$$

$$\mathbf{g}_{\partial \eta i} = \eta_{N i} \boldsymbol{\tau}, \tag{9.8.12}$$

$$\mathbf{g}_{D i} = \eta_{D i} \phi'. \tag{9.8.13}$$

In this book, the scalar product of $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{d \times d}$ and $\mathbf{B} = (b_{ij}) \in \mathbb{R}^{d \times d}$, $\sum_{(i,j) \in \{1, \dots, d\}^2} a_{ij} b_{ij}$ is written as $\mathbf{A} \cdot \mathbf{B}$. Moreover, in deriving Eq. (9.8.5), the identity

$$\mathbf{a} \cdot (\mathbf{B} \mathbf{c}) = (\mathbf{B}^\top \mathbf{a}) \cdot \mathbf{c} = (\mathbf{a} \mathbf{c}^\top) \cdot \mathbf{B} \tag{9.8.14}$$

with respect to $\mathbf{a} \in \mathbb{R}^d$, $\mathbf{B} \in \mathbb{R}^{d \times d}$ and $\mathbf{c} \in \mathbb{R}^d$ was used. Hereinafter, these relationships will be used without explanation.

Using the results above, the following results regarding \mathbf{g}_i of Eq. (9.8.5) can be obtained.

Theorem 9.8.2 (Shape Derivative \mathbf{g}_i of f_i) *Let $\phi \in \mathcal{D}$, b , p_N , u_D , ζ_i , $\eta_{N i}$ and $\eta_{D i}$ be given as functions fixed with the material satisfying Hypotheses 9.5.1, 9.5.3 and 9.6.1. Moreover, let u and v_i be the weak solutions of the state determination problem (Problem 9.5.4) and the adjoint problem (Problem 9.8.1) with respect to f_i , respectively, and are both in \mathcal{S} of Eq. (9.5.2). When $\mathbf{g}_{\partial p i}$ and $\mathbf{g}_{\partial \eta i}$ in Eqs. (9.8.10) and (9.8.12), respectively, are zero, the shape derivative of f_i becomes Eq. (9.8.5) and \mathbf{g}_i is in X' . Furthermore, we have*

$$\mathbf{G}_{\Omega i} \in H^1 \cap L^\infty (\Omega(\phi); \mathbb{R}^{d \times d}),$$

$$g_{\Omega i} \in H^1 \cap L^\infty (\Omega(\phi); \mathbb{R}),$$

$$\begin{aligned}
\mathbf{g}_{\zeta bi} &\in H^1 \cap L^\infty \left(\Omega(\boldsymbol{\phi}); \mathbb{R}^d \right), \\
\mathbf{g}_{pi} &\in H^{1/2} \cap L^\infty \left(\Gamma_p(\boldsymbol{\phi}); \mathbb{R}^d \right), \\
\mathbf{g}_{\eta i} &\in H^{1/2} \cap L^\infty \left(\Gamma_{\eta i}(\boldsymbol{\phi}); \mathbb{R}^d \right), \\
\mathbf{g}_{Di} &\in H^{1/2} \cap L^\infty \left(\Gamma_D(\boldsymbol{\phi}); \mathbb{R}^d \right).
\end{aligned}
\quad \square$$

Proof The fact that the shape derivative of f_i becomes \mathbf{g}_i of Eq. (9.8.5) is as seen above. The following holds with respect to the regularity of \mathbf{g}_i . With respect to the first term on $\mathbf{G}_{\Omega i}$, from Hölder's inequality (Theorem A.9.1) and the corollary of Poincaré inequality (Corollary A.9.4), the inequalities

$$\begin{aligned}
&\left\| \left\{ \nabla u (\nabla v_i)^\top \right\} \cdot (\nabla \boldsymbol{\varphi}^\top) \right\|_{L^1(\Omega(\boldsymbol{\phi}); \mathbb{R})} \\
&\leq \left\| \nabla u (\nabla v_i)^\top \right\|_{L^2(\Omega(\boldsymbol{\phi}); \mathbb{R}^{d \times d})} \left\| \nabla \boldsymbol{\varphi}^\top \right\|_{L^2(\Omega(\boldsymbol{\phi}); \mathbb{R}^{d \times d})} \\
&\leq \|\nabla u\|_{L^4(\Omega(\boldsymbol{\phi}); \mathbb{R}^d)} \|\nabla v_i\|_{L^4(\Omega(\boldsymbol{\phi}); \mathbb{R}^d)} \left\| \nabla \boldsymbol{\varphi}^\top \right\|_{L^2(\Omega(\boldsymbol{\phi}); \mathbb{R}^{d \times d})} \\
&\leq \|u\|_{W^{1,4}(D; \mathbb{R})} \|v_i\|_{W^{1,4}(D; \mathbb{R})} \|\boldsymbol{\varphi}\|_X \\
&\leq \|u\|_{W^{2,4}(D; \mathbb{R})} \|v_i\|_{W^{2,4}(D; \mathbb{R})} \|\boldsymbol{\varphi}\|_X
\end{aligned}$$

hold. From the assumptions, the right-hand side is finite. Hence, $\nabla u (\nabla v_i)^\top$ is in X' . Moreover, in view of the inequalities above, $\nabla u (\nabla v_i)^\top$ is also contained in $H^1 \cap L^\infty(\Omega(\boldsymbol{\phi}); \mathbb{R}^{d \times d})$. A similar result can be obtained with respect to other terms of $\mathbf{G}_{\Omega i}$. A similar result is also obtained with respect to $\mathbf{g}_{\Omega i}$. The result for $\mathbf{g}_{\zeta bi}$ is obvious from Hypotheses 9.5.1 and 9.6.1.

The regularity of \mathbf{g}_{pi} depends on the regularity of \mathbf{v} and κ in addition to regularities of v_i and p_N . With respect to the first term on the right-hand side of Eq. (9.8.9), we have

$$\begin{aligned}
&\|\kappa p_N v_i \mathbf{v} \cdot \boldsymbol{\varphi}\|_{L^1(\Gamma_p(\boldsymbol{\phi}); \mathbb{R})} \\
&\leq \|\kappa p_N v_i \mathbf{v}\|_{L^2(\Gamma_p(\boldsymbol{\phi}); \mathbb{R}^d)} \|\boldsymbol{\varphi}\|_{L^2(\Gamma_p(\boldsymbol{\phi}); \mathbb{R}^d)} \\
&\leq \|\kappa\|_{H^{1/2} \cap L^\infty(\Gamma_p(\boldsymbol{\phi}); \mathbb{R})} \|p_N\|_{L^4(\Gamma_p(\boldsymbol{\phi}); \mathbb{R})} \|v_i\|_{L^4(\Gamma_p(\boldsymbol{\phi}); \mathbb{R})} \\
&\quad \times \|\mathbf{v}\|_{H^{3/2} \cap C^{0,1}(\Gamma_p(\boldsymbol{\phi}); \mathbb{R}^d)} \|\boldsymbol{\varphi}\|_{L^2(\Gamma_p(\boldsymbol{\phi}); \mathbb{R}^d)} \\
&\leq \|\gamma_{\partial\Omega}\|^3 \|p_N\|_{W^{1,4}(\Omega(\boldsymbol{\phi}); \mathbb{R})} \|v_i\|_{W^{1,4}(\Omega(\boldsymbol{\phi}); \mathbb{R})} \\
&\quad \times \|\kappa\|_{H^{1/2} \cap L^\infty(\Gamma_p(\boldsymbol{\phi}); \mathbb{R})} \|\mathbf{v}\|_{H^{3/2} \cap C^{0,1}(\Gamma_p(\boldsymbol{\phi}); \mathbb{R}^d)} \|\boldsymbol{\varphi}\|_X \\
&\leq \|\gamma_{\partial\Omega}\|^3 \|p_N\|_{C^{1,1}(\Omega(\boldsymbol{\phi}); \mathbb{R})} \|v_i\|_{W^{2,4}(\Omega(\boldsymbol{\phi}); \mathbb{R})} \\
&\quad \times \|\kappa\|_{H^{1/2} \cap L^\infty(\Gamma_p(\boldsymbol{\phi}); \mathbb{R})} \|\mathbf{v}\|_{H^{3/2} \cap C^{0,1}(\Gamma_p(\boldsymbol{\phi}); \mathbb{R}^d)} \|\boldsymbol{\varphi}\|_X
\end{aligned}$$

using Hölder's inequality (Theorem A.9.1) and the trace theorem (Theorem 4.4.2). Here,

$$\gamma_{\partial\Omega} : H^1(\Omega(\phi); \mathbb{R}^d) \rightarrow H^{1/2}(\partial\Omega(\phi); \mathbb{R}^d)$$

is a trace operator and its operator norm $\|\gamma_{\partial\Omega}\|$ is bounded from the fact that the boundary $\partial\Omega(\phi)$ is Lipschitz. Moreover, $\Gamma_p(\phi)$ is defined to be piecewise $H^3 \cap C^{1,1}$ in \mathcal{D} of Eq. (9.1.3). Hence, v is in the class of $H^{3/2} \cap C^{0,1}$ and κ is in the class of $H^{1/2} \cap L^\infty$ on $\Gamma_p(\phi)$. Therefore, $\kappa p_N v_i v$ is an element of X' and is in $H^{1/2} \cap H^{1/2}(\Gamma_p(\phi); \mathbb{R}^d) \cap L^\infty(\Gamma_p(\phi); \mathbb{R}^d)$. The second term on the right-hand side of Eq. (9.8.9) belongs to $H^{1/2} \cap L^\infty(\Gamma_p(\phi); \mathbb{R}^d)$ because $\tau_1(\phi), \dots, \tau_{d-1}(\phi)$ is in the class $H^{3/2} \cap C^{0,1}$, $p_N \in C^{1,1}(D; \mathbb{R})$ (Hypothesis 9.5.1) and $v_i \in W^{2,4}(D; \mathbb{R})$ on $\Gamma_p(\phi)$ (Practice 9.1). The third term of Eq. (9.8.9) becomes $v_i p'_N \in W^{2,4}(D; \mathbb{R})$. Hence, $g_{pi} \in H^{1/2} \cap L^\infty(\Gamma_p(\phi); \mathbb{R}^d)$ is shown.

For the regularities of g_{η_i} , the same result as for g_{pi} can be obtained from $\nabla \eta_{Ni} \in W^{1,q_R}(D; \mathbb{R})$. Moreover, the result for g_{Di} is obvious from Hypotheses 9.6.1. Therefore, the result of the theorem is established.

In addition, we assumed $g_{\partial pi} = \mathbf{0}_{\mathbb{R}^d}$ because the trace of $\varphi \in X$ on $\partial\Gamma_p(\phi) \cup \Theta_p(\phi)$ can not be defined. Similarly, $g_{\partial\eta_i} = \mathbf{0}_{\mathbb{R}^d}$ was assumed. \square

9.8.2 Second-Order Shape Derivative of f_i Using Formulae Based on Shape Derivative of a Function

Let us obtain the second-order shape derivative of the cost function based on the method shown in Sect. 7.5.3. Here, the formulae using the shape derivative of a function is used.

In order to obtain the second-order shape derivative of \tilde{f}_i , the following assumptions are established.

Hypothesis 9.8.3 (Second-Order Shape Derivative of \tilde{f}_i) With respect to the state determination problem (Problem 9.5.4) and the cost function f_i defined in Eq. (9.6.1), the following assumptions are made, respectively:

- (1) $b = 0$, $\zeta_{i\phi'}(\phi, u, \nabla u)[\varphi] = 0$.
- (2) ζ_i is not a function of ϕ and u , but is a bilinear form of ∇u .
- (3) Equations (9.8.9) to (9.8.13) are zero, or $\tilde{\Gamma}_0 = \Gamma_{p0} \cup \Gamma_{\eta i 0} \in \bar{\Omega}_{C0}$ in Eq. (9.1.1).

The Lagrange function \mathcal{L}_i of f_i is defined by Eq. (9.6.3). Viewing (ϕ, u) as a design variable and putting its admissible set and admissible direction set as

$$S = \{(\phi, u) \in \mathcal{D} \times \mathcal{S} \mid \mathcal{L}_S(\phi, u, v) = 0 \text{ for all } v \in U\},$$

$$T_S(\phi, u) = \{(\varphi, \hat{v}) \in X \times U \mid \mathcal{L}_{S\phi u}(\phi, u, v)[\varphi, \hat{v}] = 0 \text{ for all } v \in U\},$$

the second-order Fréchet partial derivative of \mathcal{L}_i with respect to arbitrary variations $(\varphi_1, \hat{v}_1), (\varphi_2, \hat{v}_2) \in T_S(\phi, u)$ of $(\phi, u) \in S$, similarly to Eq. (7.5.21), and considering Eq. (9.1.6), becomes

$$\begin{aligned}
& \mathcal{L}_{i(\phi', u)(\phi', u)}(\phi, u, v_i)[(\varphi_1, \hat{v}_1), (\varphi_2, \hat{v}_2)] \\
&= (\mathcal{L}_{0(\phi', u)})_{(\phi', u)}(\phi, u, v_i)[(\varphi_1, \hat{v}_1), (\varphi_2, \hat{v}_2)] + \langle g_0(\phi), t(\varphi_1, \varphi_2) \rangle \\
&= (\mathcal{L}_{i\phi'}(\phi, u, v_i)[\varphi_1] + \mathcal{L}_{iu}(\phi, u, v_i)[\hat{v}_1])_{\phi'}[\varphi_2] \\
&\quad + (\mathcal{L}_{i\phi'}(\phi, u, v_i)[\varphi_1] + \mathcal{L}_{iu}(\phi, u, v_i)[\hat{v}_1])_u[\hat{v}_2] \\
&\quad + \langle g_0(\phi), t(\varphi_1, \varphi_2) \rangle \\
&= (\mathcal{L}_{i\phi'}(\phi, u, v_i)[\varphi_1, \varphi_2] + \mathcal{L}_{i\phi'u}(\phi, u, v_i)[\varphi_1, \hat{v}_2] \\
&\quad + \mathcal{L}_{i\phi'u}(\phi, u, v_i)[\varphi_2, \hat{v}_1] + \mathcal{L}_{iuu}(\phi, u, v_i)[\hat{v}_1, \hat{v}_2] \\
&\quad + \langle g_0(\phi), t(\varphi_1, \varphi_2) \rangle), \tag{9.8.15}
\end{aligned}$$

where $\langle g_0(\phi), t(\varphi_1, \varphi_2) \rangle$ follows the definition given in Eq. (9.1.8) (or Eq. (9.3.10)).

The first and fifth terms on the right-hand side of Eq. (9.8.15) become

$$\begin{aligned}
& (\mathcal{L}_{i\phi'}(\phi, u, v_i)[\varphi_1, \varphi_2] + \langle g_0(\phi), t(\varphi_1, \varphi_2) \rangle) \\
&= \int_{\Omega(\phi)} \left[\left\{ \nabla u \cdot (\nabla \varphi_1^\top \nabla v_i) \right\}_{\phi'}[\varphi_2] \right. \\
&\quad + \left\{ \nabla v_i \cdot (\nabla \varphi_1^\top \nabla u) \right\}_{\phi'}[\varphi_2] - \left\{ \zeta_{i(\nabla u)^\top} \cdot (\nabla \varphi_1^\top \nabla u) \right\}_{\phi'}[\varphi_2] \\
&\quad + (\zeta_i - \nabla u \cdot \nabla v_i) (\nabla \cdot \varphi_1)_{\phi'}[\varphi_2] \\
&\quad + \left\{ \nabla u (\nabla v_i)^\top + \nabla v_i (\nabla u)^\top - \zeta_{i(\nabla u)^\top} (\nabla u)^\top \right\} \\
&\quad \cdot \left\{ \nabla \varphi_2^\top \nabla \varphi_1^\top - \nabla \varphi_1^\top (\nabla \cdot \varphi_2) \right\} \\
&\quad \left. + (\zeta_i - \nabla u \cdot \nabla v_i) \left\{ (\nabla \varphi_2^\top)^\top \cdot \nabla \varphi_1^\top - (\nabla \cdot \varphi_2) (\nabla \cdot \varphi_1) \right\} \right] dx, \tag{9.8.16}
\end{aligned}$$

by using the first term on the right-hand side of Eqs. (9.8.4) and (9.3.11). The first integrand on the right-hand side of Eq. (9.8.16) can be expressed as follows:

$$\left\{ \nabla u \cdot (\nabla \varphi_1^\top \nabla v_i) \right\}_{\phi'}[\varphi_2]$$

$$\begin{aligned}
&= - \left\{ \nabla u \cdot (\nabla \varphi_1^\top \nabla v_i) \right\}_{\nabla u} \cdot (\nabla \varphi_2^\top \nabla u) \\
&\quad - \left\{ \nabla u \cdot (\nabla \varphi_1^\top \nabla v_i) \right\}_{\nabla \varphi_1^\top} \cdot (\nabla \varphi_2^\top \nabla \varphi_1^\top) \\
&\quad - \left\{ \nabla u \cdot (\nabla \varphi_1^\top \nabla v_i) \right\}_{\nabla v_i} \cdot (\nabla \varphi_2^\top \nabla v_i) \\
&\quad + \left\{ \nabla u \cdot (\nabla \varphi_1^\top \nabla v_i) \right\} (\nabla \cdot \varphi_2) \\
&= - (\nabla \varphi_2^\top \nabla u) \cdot (\nabla \varphi_1^\top \nabla v_i) - \nabla u \cdot (\nabla \varphi_2^\top \nabla \varphi_1^\top \nabla v_i) \\
&\quad - \nabla u \cdot (\nabla \varphi_1^\top \nabla \varphi_2^\top \nabla v_i) + \left\{ \nabla u \cdot (\nabla \varphi_1^\top \nabla v_i) \right\} \nabla \cdot \varphi_2 \\
&= - \left\{ \nabla u (\nabla v_i)^\top \right\} \cdot \left\{ (\nabla \varphi_2^\top)^\top \nabla \varphi_1^\top \right\} - \left\{ \nabla u (\nabla v_i)^\top \right\} \cdot (\nabla \varphi_2^\top \nabla \varphi_1^\top) \\
&\quad - \left\{ \nabla u (\nabla v_i)^\top \right\} \cdot (\nabla \varphi_1^\top \nabla \varphi_2^\top) + \left\{ \nabla u (\nabla v_i)^\top \right\} \cdot \nabla \varphi_1^\top (\nabla \cdot \varphi_2).
\end{aligned} \tag{9.8.17}$$

In Eq. (9.8.17), the identities in Eq. (9.8.14) and

$$\begin{aligned}
\mathbf{A} \cdot (\mathbf{B} \mathbf{C}) &= (\mathbf{B}^\top \mathbf{A}) \cdot \mathbf{C} = (\mathbf{A} \mathbf{C}^\top) \cdot \mathbf{B}, \\
(\mathbf{A} \mathbf{B}) \cdot \mathbf{C} &= \mathbf{B} \cdot (\mathbf{A}^\top \mathbf{C}) = \mathbf{A} \cdot (\mathbf{C} \mathbf{B}^\top),
\end{aligned} \tag{9.8.18}$$

where $\mathbf{A} \in \mathbb{R}^{d \times d}$, $\mathbf{B} \in \mathbb{R}^{d \times d}$ and $\mathbf{C} \in \mathbb{R}^{d \times d}$, were used. In the remaining part, those identities will be used frequently.

Similarly, the second integrand on the right-hand side of Eq. (9.8.16) is similar to Eq. (9.8.17) with u and v_i interchanged. Meanwhile, the third integrand on the right-hand side of Eq. (9.8.16) becomes similar to Eq. (9.8.17) with u and v_i interchanged and v_i and $\zeta_{i(\nabla u)^\top}$ interchanged. Lastly, the fourth integrand on the right-hand side of Eq. (9.8.16) becomes

$$\begin{aligned}
&(\zeta_i - \nabla u \cdot \nabla v_i) (\nabla \cdot \varphi_1)_{\varphi'} [\varphi_2] \\
&= (\zeta_i - \nabla u \cdot \nabla v_i) \left\{ - (\nabla \varphi_2^\top)^\top \cdot \nabla \varphi_1^\top + (\nabla \cdot \varphi_2) (\nabla \cdot \varphi_1) \right\}.
\end{aligned}$$

Hence, Eq. (9.8.16) becomes

$$\begin{aligned}
&(\mathcal{L}_{i\varphi'})_{\varphi'} (\varphi, u, v_i) [\varphi_1, \varphi_2] + \langle \mathbf{g}_0(\varphi), \mathbf{t}(\varphi_1, \varphi_2) \rangle \\
&= \int_{\Omega(\varphi)} \left[- \left\{ \nabla u (\nabla v_i)^\top + \nabla v_i (\nabla u)^\top - \zeta_{i(\nabla u)^\top} (\nabla u)^\top \right\} \right. \\
&\quad \left. \cdot \left\{ \nabla \varphi_1^\top \nabla \varphi_2^\top + (\nabla \varphi_2^\top)^\top \nabla \varphi_1^\top \right\} \right] dx.
\end{aligned} \tag{9.8.19}$$

Next, we look at the second term on the right-hand side of Eq. (9.8.15). If the first term on the right-hand side of Eq. (9.8.4) is used, we get

$$\begin{aligned} & \mathcal{L}_{i\phi'u}(\phi, u, v_i)[\varphi_1, \hat{v}_2] \\ &= \int_{\Omega(\phi)} \left\{ \nabla \hat{v}_2 \cdot (\nabla \varphi_1^\top \nabla v_i) + (\nabla v_i - \zeta_{i(\nabla u)^\top}) \cdot (\nabla \varphi_1^\top \nabla \hat{v}_2) \right. \\ &\quad \left. - (\nabla \hat{v}_2 \cdot \nabla v_i) \nabla \cdot \varphi_1 \right\} dx. \end{aligned} \quad (9.8.20)$$

On the other hand, the variation of u satisfying the state determination problem with respect to an arbitrary domain variation $\varphi_j \in Y$ for $j \in \{1, 2\}$ is given as $\hat{v}_j = v'(\phi)[\varphi_j]$. If the Fréchet partial derivative of the Lagrange function \mathcal{L}_S of the state determination problem defined by Eq. (9.5.3) is taken, we obtain

$$\begin{aligned} & \mathcal{L}_{S\phi'u}(\phi, u, v)[\varphi_j, \hat{v}_j] \\ &= \int_{\Omega(\phi)} \left\{ \nabla u \cdot (\nabla \varphi_j^\top \nabla v) + \nabla v \cdot (\nabla \varphi_j^\top \nabla u) \right. \\ &\quad \left. - (\nabla u \cdot \nabla v) \nabla \cdot \varphi_j - \nabla \hat{v}_j \cdot \nabla v \right\} dx \\ &= \int_{\Omega(\phi)} \left[\left\{ \left((\nabla \varphi_j^\top)^\top + \nabla \varphi_j^\top - \nabla \cdot \varphi_j \right) \nabla u - \nabla \hat{v}_j \right\} \cdot \nabla v \right] dx \\ &= 0 \end{aligned} \quad (9.8.21)$$

for all $v \in U$. Here, Hypothesis 9.8.3 and the fact that v and \hat{v}_j are both zero on Γ_D were used. From Eq. (9.8.21), we get

$$\nabla \hat{v}_j = \left\{ (\nabla \varphi_j^\top)^\top + \nabla \varphi_j^\top - \nabla \cdot \varphi_j \right\} \nabla u. \quad (9.8.22)$$

This relation becomes possible by the following argument. Substituting Eq. (9.8.22) into Eq. (9.8.20), the second term on the right-hand side of Eq. (9.8.15) becomes

$$\begin{aligned} & \mathcal{L}_{i\phi'u}(\phi, u, v_i)[\varphi_1, \hat{v}_2] \\ &= \int_{\Omega(\phi)} \left[\left\{ \left((\nabla \varphi_2^\top)^\top + \nabla \varphi_2^\top - \nabla \cdot \varphi_2 \right) \nabla u (\nabla v_i)^\top \right. \right. \\ &\quad \left. \left. + (\nabla v_i - \zeta_{i(\nabla u)^\top}) (\nabla u)^\top \left((\nabla \varphi_2^\top)^\top + \nabla \varphi_2^\top - \nabla \cdot \varphi_2 \right) \right\} \cdot \nabla \varphi_1^\top \right. \\ &\quad \left. - \left\{ \left((\nabla \varphi_2^\top)^\top + \nabla \varphi_2^\top - \nabla \cdot \varphi_2 \right) \nabla u \right\} \cdot \nabla v_i \right] \nabla \cdot \varphi_1 dx \end{aligned}$$

$$\begin{aligned}
&= \int_{\Omega(\phi)} \left[\left\{ \nabla u (\nabla v_i)^\top + \nabla v_i (\nabla u)^\top - \zeta_{i(\nabla u)^\top} (\nabla u)^\top \right\} \right. \\
&\quad \cdot \left\{ \nabla \varphi_1^\top \nabla \varphi_2^\top + \nabla \varphi_1^\top (\nabla \varphi_2^\top)^\top - \nabla \varphi_1^\top (\nabla \cdot \varphi_2) \right\} \\
&\quad - \left\{ \nabla u (\nabla v_i)^\top \right\} \cdot \left\{ \nabla \varphi_2^\top (\nabla \cdot \varphi_1) + (\nabla \varphi_2^\top)^\top (\nabla \cdot \varphi_1) \right\} \\
&\quad \left. + \nabla u \cdot \nabla v_i (\nabla \cdot \varphi_1) (\nabla \cdot \varphi_2) \right] dx. \tag{9.8.23}
\end{aligned}$$

Similarly, the third term on the right-hand side of Eq. (9.8.15) becomes $\mathcal{L}_{iu\phi'}(\phi, u, v_i)[\varphi_2, \nabla \hat{v}_1]$ similar to Eq. (9.8.23) with φ_1 and φ_2 interchanged. Lastly, the fourth term on the right-hand side of Eq. (9.8.15) vanishes.

Summarizing the results above, the second-order shape derivative of \tilde{f}_i becomes

$$\begin{aligned}
h_i(\phi, u, u)[\varphi_1, \varphi_2] &= \int_{\Omega(\phi)} \left[2 \nabla u \cdot \nabla v_i(u) (\nabla \cdot \varphi_2) (\nabla \cdot \varphi_1) \right. \\
&\quad + \left\{ \nabla u (\nabla v_i)^\top + \nabla v_i (\nabla u)^\top - \zeta_{i(\nabla u)^\top} (\nabla u)^\top \right\} \\
&\quad \cdot \left\{ \nabla \varphi_1^\top (\nabla \varphi_2^\top)^\top + \nabla \varphi_2^\top \nabla \varphi_1^\top - \nabla \varphi_1^\top \nabla \cdot \varphi_2 - \nabla \varphi_2^\top (\nabla \cdot \varphi_1) \right\} \\
&\quad - \left\{ \nabla u (\nabla v_i)^\top \right\} \\
&\quad \cdot \left\{ \nabla \varphi_2^\top (\nabla \cdot \varphi_1) + (\nabla \varphi_2^\top)^\top (\nabla \cdot \varphi_1) \right. \\
&\quad \left. + \nabla \varphi_1^\top (\nabla \cdot \varphi_2) + (\nabla \varphi_1^\top)^\top (\nabla \cdot \varphi_2) \right\} \right] dx. \tag{9.8.24}
\end{aligned}$$

9.8.3 Second-Order Shape Derivative of Cost Function Using Lagrange Multiplier Method

When the Lagrange multiplier method is used to obtain the second-order shape derivative of a cost function, we use the same idea given in Sect. 7.5.4. Fixing φ_1 , we define the Lagrange function with respect to $\tilde{f}'_i(\phi)[\varphi_1] = \langle \mathbf{g}_i, \varphi_1 \rangle$ in Eq. (9.8.5) by

$$\mathcal{L}_{li}(\phi, u, v_i, w_i, z_i) = \langle \mathbf{g}_i, \varphi_1 \rangle + \mathcal{L}_S(\phi, u, w_i) + \mathcal{L}_{Ai}(\phi, v_i, z_i), \tag{9.8.25}$$

where \mathcal{L}_S is given by Eq. (9.5.3), and

$$\begin{aligned} \mathcal{L}_{Ai}(\phi, v_i, z_i) &= \int_{\Omega(\phi)} (-\nabla v_i \cdot \nabla z_i + \zeta_{iu} z_i + \zeta_{i(\nabla u)^\top} \cdot \nabla z_i) dx \\ &+ \int_{\Gamma_{\eta i}(\phi)} \eta_{Niu} z_i d\gamma + \int_{\Gamma_D(\phi)} \{z_i \partial_\nu v_i + (v_i - \eta_{Di} \partial_\nu u) \partial_\nu z_i\} d\gamma \end{aligned} \quad (9.8.26)$$

is the Lagrange function with respect to the adjoint problem (Problem 9.8.1). $w_i \in U$ and $z_i \in U$ are the adjoint variables provided for u and v_i in \mathbf{g}_i .

With respect to arbitrary variations $(\varphi_2, \hat{u}, \hat{v}_i, \hat{w}_i, \hat{z}_i) \in \mathcal{D} \times U^4$ of (ϕ, u, v_i, w_i, z_i) , considering Eq. (9.1.6), the Fréchet derivative of \mathcal{L}_{li} is written as

$$\begin{aligned} \mathcal{L}'_{li}(\phi, u, v_i, w_i, z_i) &[\varphi_2, \hat{u}, \hat{v}_i, \hat{w}_i, \hat{z}_i] \\ &= \mathcal{L}_{li\phi'}(\phi, u, v_i, w_i, z_i) [\varphi_2] + \langle \mathbf{g}_0(\phi), \mathbf{t}(\varphi_1, \varphi_2) \rangle \\ &+ \mathcal{L}_{liu}(\phi, u, v_i, w_i, z_i) [\hat{u}] + \mathcal{L}_{liv_i}(\phi, u, v_i, w_i, z_i) [\hat{v}_i] \\ &+ \mathcal{L}_{liw_i}(\phi, u, v_i, w_i, z_i) [\hat{w}_i] + \mathcal{L}_{liz_i}(\phi, u, v_i, w_i, z_i) [\hat{z}_i]. \end{aligned} \quad (9.8.27)$$

The fifth term on the right-hand side of Eq. (9.8.27) vanishes if u is the solution of the state determination problem. If v_i can be determined as the solution of the adjoint problem, the sixth term of Eq. (9.8.27) also vanishes.

Applying Proposition 9.3.7, the third term on the right-hand side of Eq. (9.8.27) is obtained as

$$\begin{aligned} \mathcal{L}_{liu}(\phi, u, v_i, w_i, z_i) &[\hat{u}] \\ &= \int_{\Omega(\phi)} \left[\left\{ \left(\nabla \varphi_1^\top + (\nabla \varphi_1^\top)^\top \right) \nabla v_i - (\nabla \varphi_1^\top)^\top \zeta_{i(\nabla u)^\top} \right. \right. \\ &\quad \left. \left. - (\nabla \cdot \varphi_1) \nabla v_i \right\} \cdot \nabla \hat{u} \right. \\ &\quad \left. - \left\{ \left((\nabla \varphi^\top)^\top \zeta_{i(\nabla u)^\top} u \right) \cdot \nabla u + (\nabla \cdot \varphi_1) \zeta_{iu} \right\} \hat{u} - \nabla w_i \cdot \nabla \hat{u} \right] dx \\ &+ \int_{\Gamma_{\eta i}(\phi)} \eta_{Niu} (\nabla \cdot \varphi_1)_\tau \hat{u} d\gamma. \end{aligned} \quad (9.8.28)$$

Here, the condition that Eq. (9.8.28) is zero for arbitrary $\hat{u} \in U$ is equivalent to setting w_i to be the solution of the following adjoint problem.

Problem 9.8.4 (Adjoint Problem of w_i with Respect to $\langle g_i, \varphi_1 \rangle$) Under the assumption of Problem 9.6.3, letting $\varphi_1 \in Y$ be given, find $w_i = w_i(\varphi_1) \in U$ satisfying

$$\begin{aligned} -\Delta w_i &= -\nabla^\top \left\{ \left(\nabla \varphi_1^\top + (\nabla \varphi_1^\top)^\top \right) \nabla v_i - (\nabla \varphi_1^\top)^\top \zeta_{i(\nabla u)^\top} \right. \\ &\quad \left. - (\nabla \cdot \varphi_1) \nabla v_i \right\} \\ &\quad - \left((\nabla \varphi^\top)^\top \zeta_{i(\nabla u)^\top} u \right) \cdot \nabla u - (\nabla \cdot \varphi_1) \zeta_{iu} \quad \text{in } \Omega(\phi), \\ \partial_\nu w_i &= \eta_{Niu} (\nabla \cdot \varphi_1)_\tau + \left\{ \left(\nabla \varphi_1^\top + (\nabla \varphi_1^\top)^\top \right) \nabla v_i - (\nabla \varphi_1^\top)^\top \zeta_{i(\nabla u)^\top} \right. \\ &\quad \left. - (\nabla \cdot \varphi_1) \nabla v_i \right\} \cdot \nu \quad \text{on } \Gamma_{\eta i}(\phi), \\ \partial_\nu w_i &= 0 \quad \text{on } \Gamma_N(\phi) \setminus \bar{\Gamma}_{\eta i}(\phi), \\ w_i &= 0 \quad \text{on } \Gamma_D(\phi). \end{aligned}$$

□

The fourth term on the right-hand side of Eq. (9.8.27) is

$$\begin{aligned} \mathcal{L}_{liv_i}(\phi, u, v_i, w_i, z_i) [\hat{v}_i] &= \int_{\Omega(\phi)} \left[\left\{ \left(\nabla \varphi_1^\top + (\nabla \varphi_1^\top)^\top \right) \nabla u - (\nabla \cdot \varphi_1) \nabla u \right\} \cdot \nabla \hat{v}_i \right. \\ &\quad \left. + b \hat{v}_i - \nabla z_i \cdot \nabla \hat{v}_i \right] dx + \int_{\Gamma_p(\phi)} p_N (\nabla \cdot \varphi_1)_\tau \hat{v}_i \, d\gamma, \end{aligned} \quad (9.8.29)$$

where φ_1 is assumed to be an H^2 class function in the neighborhood of $\Gamma_{\eta i}(\phi)$. Here, the condition that Eq. (9.8.29) is zero for arbitrary $\hat{v}_i \in U$ is equivalent to setting z_i to be the solution of the following adjoint problem.

Problem 9.8.5 (Adjoint Problem of z_i with Respect to $\langle g_i, \varphi_1 \rangle$) Under the assumption of Problem 9.6.3, letting $\varphi_1 \in Y$ be given, find $z_i = z_i(\varphi_1) \in U$ satisfying

$$\begin{aligned} -\Delta z_i &= b - \nabla^\top \left\{ \left(\nabla \varphi_1^\top + (\nabla \varphi_1^\top)^\top \right) \nabla u - (\nabla \cdot \varphi_1) \nabla u \right\} \quad \text{in } \Omega(\phi), \\ \partial_\nu z_i &= p_N (\nabla \cdot \varphi_1)_\tau \\ &\quad + \left\{ \left(\nabla \varphi_1^\top + (\nabla \varphi_1^\top)^\top \right) \nabla u - (\nabla \cdot \varphi_1) \nabla u \right\} \cdot \nu \quad \text{on } \Gamma_p(\phi), \\ \partial_\nu z_i &= 0 \quad \text{on } \Gamma_N(\phi) \setminus \bar{\Gamma}_p(\phi), \\ z_i &= 0 \quad \text{on } \Gamma_D. \end{aligned}$$

□

Finally, the first and second terms on the right-hand side of Eq. (9.8.27) become

$$\begin{aligned} & \mathcal{L}_{\text{li}\phi'}(\phi, u, v_i, w_i(\varphi_1), z_i(\varphi_1))[\varphi_2] + \langle g_0(\phi), t(\varphi_1, \varphi_2) \rangle \\ &= \mathcal{L}_{i\phi'\phi'}(\phi, u, v_i)[\varphi_1, \varphi_2] + \langle g_0(\phi), t(\varphi_1, \varphi_2) \rangle \\ & \quad + \mathcal{L}_{S\phi'}(\phi, u, w_i)[\varphi_2] + \mathcal{L}_{Ai\phi'}(\phi, v_i, z_i)[\varphi_2]. \end{aligned} \quad (9.8.30)$$

The first and second terms of Eq. (9.8.30) are given by Eq. (9.8.19). The third and fourth terms become

$$\begin{aligned} & \mathcal{L}_{S\phi'}(\phi, u, w_i)[\varphi_2] \\ &= \int_{\Omega(\phi)} \left[\nabla u \cdot \left\{ (\nabla \varphi_2^\top) \nabla w_i(\varphi_1) \right\} + \nabla w_i(\varphi_1) \cdot \left\{ (\nabla \varphi_2^\top) \nabla u \right\} \right. \\ & \quad \left. + (bw_i(\varphi_1) - \nabla u \cdot \nabla w_i(\varphi_1)) \nabla \cdot \varphi_2 \right] dx \\ & \quad + \int_{\Gamma_{\eta i}(\phi)} \eta_{Ni} (\nabla \cdot \varphi_1)_\tau (\nabla \cdot \varphi_2)_\tau d\gamma \\ & \quad + \int_{\Gamma_p(\phi)} p_N w_i(\varphi_1) (\nabla \cdot \varphi_2)_\tau d\gamma, \\ & \mathcal{L}_{Ai\phi'}(\phi, v_i, z_i)[\varphi_2] \\ &= \int_{\Omega(\phi)} \left[(\nabla v_i - \xi_{i(\nabla u)^\top}) \cdot \left\{ (\nabla \varphi_2^\top) \nabla z_i(\varphi_1) \right\} \right. \\ & \quad \left. + \nabla z_i(\varphi_1) \cdot \left\{ (\nabla \varphi_2^\top) \nabla v_i \right\} \right. \\ & \quad \left. + (\xi_{iu} z_i(\varphi_1) - \nabla u \cdot \nabla z_i(\varphi_1)) \nabla \cdot \varphi_2 \right] dx \\ & \quad + \int_{\Gamma_{\eta i}(\phi)} \eta_{Ni u} z_i(\varphi_1) (\nabla \cdot \varphi_2)_\tau d\gamma \\ & \quad + \int_{\Gamma_p(\phi)} p_N v_i(\nabla \cdot \varphi_1)_\tau (\nabla \cdot \varphi_2)_\tau d\gamma \end{aligned}$$

with respect to an arbitrary variation $\varphi_1 \in Y$.

Here, u , v_i , $w_i(\varphi_1)$ and $z_i(\varphi_1)$ are assumed to be the weak solutions of Problems 9.5.4, 9.8.1, 9.8.4 and 9.8.5, respectively. If we denote $f_i(\phi, u)$ by $\tilde{f}_i(\phi)$, then we obtain the relation

$$\begin{aligned} & \mathcal{L}_{\text{li}\phi'}(\phi, u, v_i, w_i(\varphi_1), z_i(\varphi_1))[\varphi_2] + \langle g_0(\phi), t(\varphi_1, \varphi_2) \rangle \\ &= \tilde{f}_i''(\phi)[\varphi_1, \varphi_2] = \langle g_{Hi}(\phi, \varphi_1), \varphi_2 \rangle \end{aligned}$$

$$\begin{aligned}
&= \int_{\Omega(\phi)} \left[- \left\{ \nabla u (\nabla v_i)^\top + \nabla v_i (\nabla u)^\top - \zeta_{i(\nabla u)^\top} (\nabla u)^\top \right\} \right. \\
&\quad \cdot \left\{ \nabla \varphi_1^\top \nabla \varphi_2^\top + (\nabla \varphi_2^\top)^\top \nabla \varphi_1^\top \right\} \\
&\quad + \left\{ \nabla u (\nabla w_i(\varphi_1))^\top + \nabla w_i(\varphi_1) (\nabla u)^\top \right. \\
&\quad \left. + (\nabla v_i - \zeta_{i(\nabla u)^\top}) (\nabla z_i(\varphi_1))^\top \right\} \cdot \nabla \varphi_2^\top \\
&\quad + (b w_i(\varphi_1) + \zeta_{i u} z_i(\varphi_1) - \nabla u \cdot \nabla w_i(\varphi_1) - \nabla u \cdot \nabla z_i(\varphi_1)) \nabla \cdot \varphi_2 \\
&\quad \left. + \zeta_{i \phi' \phi'}(\phi, u, \nabla u) [\varphi_1, \varphi_2] + u b''(\phi) [\varphi_1, \varphi_2] \right] dx \\
&+ \int_{\Gamma_{\eta_i}(\phi)} \left[\{ \eta_{Ni} (\nabla \cdot \varphi_1)_\tau + \eta_{Ni u} z_i(\varphi_1) \} (\nabla \cdot \varphi_2)_\tau \right. \\
&\quad \left. + \eta_{Ni \phi' \phi'}(\phi, u) [\varphi_1, \varphi_2] \right] d\gamma \\
&+ \int_{\Gamma_p(\phi)} [\{ p_N v_i (\nabla \cdot \varphi_1)_\tau + p_N w_i(\varphi_1) \} (\nabla \cdot \varphi_2)_\tau + p''(\phi) [\varphi_1, \varphi_2]] d\gamma \\
&+ \int_{\Gamma_D(\phi)} \eta_{Di \phi' \phi'}(\phi, \partial_\nu u) [\varphi_1, \varphi_2] d\gamma, \tag{9.8.31}
\end{aligned}$$

where $\mathbf{g}_{\text{Hi}}(\phi, \varphi_1)$ is the Hesse gradient of f_i .

9.8.4 Shape Derivative of f_i Using Formulae Based on Partial Shape Derivative of a Function

Next, we compute the shape derivative of f_i by computing the shape derivative of \mathcal{L}_i using the formulae for the partial shape derivative of a function shown in Sect. 9.3.2.

Here, we assume that u and v_i are elements such that $u - u_D$ and $v_i - \eta_{Di \partial_\nu u}$ belong to $U(\phi) \cap W^{2,2q_R}(D; \mathbb{R})$ ($q_R > d$). Hypotheses 9.5.2 and 9.6.2 give the conditions for these.

Under these assumptions, the Fréchet derivative of $\mathcal{L}_i(\phi, u, v_i)$, with respect to an arbitrary $(\phi, \hat{u}, \hat{v}_i) \in X \times U \times U$ can be written as

$$\begin{aligned}
&\mathcal{L}'_i(\phi, u, v_i) [\varphi, \hat{u}, \hat{v}_i] \\
&= \mathcal{L}_{i\phi^*}(\phi, u, v_i) [\varphi] + \mathcal{L}_{iu}(\phi, u, v_i) [\hat{u}] + \mathcal{L}_{iv_i}(\phi, u, v_i) [\hat{v}_i] \tag{9.8.32}
\end{aligned}$$

using the notation of Eqs. (9.3.21) and (9.3.27). Unlike Sect. 9.8.1, since Eqs. (9.3.21) and (9.3.27) were used here, u^* was replaced by arbitrary $\hat{u} \in X$. Let us look at the detail of each term below.

The last term on the right-hand side of Eq. (9.8.32) becomes

$$\mathcal{L}_{i v_i}(\phi, u, v_i)[\hat{v}_i] = \mathcal{L}_{S v_i}(\phi, u, v_i)[\hat{v}_i] = \mathcal{L}_S(\phi, u, \hat{v}_i). \quad (9.8.33)$$

Equation (9.8.33) is the Lagrange function of the state determination problem (Problem 9.5.4). Hence, if u is the weak solution of the state determination problem, this expression vanishes.

Meanwhile, the second term on the right-hand side of Eq. (9.8.32) is the same as Eq. (9.8.3). Hence, if v_i is such that Eq. (9.8.3) is zero with respect to an arbitrary $\hat{u} \in U$, the second term on the right-hand side of Eq. (9.8.32) also vanishes. This relationship holds when v_i is the weak solution of an adjoint problem (Problem 9.8.1) with respect to f_i . The regularities for ξ_{iu} , $\xi_{i(\nabla u)^\top}$, $\eta_{Ni u}$, $\eta_{Di \partial_v u}$ in Hypothesis 9.6.2 give the conditions that $v_i - \eta_{Di \partial_v u}$ belongs to $U(\phi) \cap W^{2,2q_R}(D; \mathbb{R})$ ($q_R > d$).

Lastly, the first term on the right-hand side of Eq. (9.8.32) is manipulated as follows. Applying the formulae of Eq. (9.3.21), representing the result of Proposition 9.3.10, and Eq. (9.3.27), representing the result of Proposition 9.3.13, to the first term, we have

$$\begin{aligned} & \mathcal{L}_{i \phi^*}(\phi, u, v_i)[\varphi] \\ &= \int_{\Omega(\phi)} (\xi_{i \phi^*} + u b^*) \cdot \varphi \, dx \\ &+ \int_{\partial \Omega(\phi)} (\xi_i(u, \nabla u) - \nabla u \cdot \nabla v_i + b v_i) \, \mathbf{v} \cdot \varphi \, d\gamma \\ &+ \int_{\Gamma_{\eta_i}(\phi)} \{(\partial_v + \kappa) \eta_{Ni}(u) \, \mathbf{v} \cdot \varphi + \eta_{Ni \phi^*} \cdot \varphi\} \, d\gamma \\ &+ \int_{\partial \Gamma_{\eta_i}(\phi) \cup \Theta_{\eta_i}(\phi)} \eta_{Ni}(u) \, \boldsymbol{\tau} \cdot \varphi \, d\varsigma \\ &+ \int_{\Gamma_p(\phi)} \{(\partial_v + \kappa) (p_N v_i) \, \mathbf{v} \cdot \varphi + v_i p_N^* \cdot \varphi\} \, d\gamma \\ &+ \int_{\partial \Gamma_p(\phi) \cup \Theta_p(\phi)} p_N v_i \, \boldsymbol{\tau} \cdot \varphi \, d\varsigma \\ &+ \int_{\Gamma_D(\phi)} \left[\{(u - u_D) \bar{w}(\varphi, v_i) + (v_i - \eta_{Di \partial_v u}) \bar{w}(\varphi, u)\} \right. \\ &\quad \left. + (\partial_v + \kappa) \{(u - u_D) \partial_v v_i + v_i \partial_v u - \eta_{Di}\} \, \mathbf{v} \cdot \varphi + \eta_{Di \phi^*} \cdot \varphi \right] \, d\gamma \\ &+ \int_{\partial \Gamma_D(\phi) \cup \Theta_D} \{(u - u_D) \partial_v v_i + v_i \partial_v u - \eta_{Di}\} \, \mathbf{v} \cdot \varphi \, d\varsigma, \end{aligned} \quad (9.8.34)$$

where $\bar{w}(\varphi, u)$ and $(\nabla \cdot \varphi)_\tau$ obey Eqs. (9.3.24) and (9.2.6), respectively.

With the above results in mind, we assume that u and v_i are the weak solutions to Problem 9.5.4 and Problem 9.8.1, respectively. In addition, we also assume that the condition for η_{Di} in Hypothesis 9.6.2 holds. Then, the notation in Eq. (7.5.15) for \tilde{f}_i can be used to write

$$\begin{aligned}\tilde{f}'_i(\phi)[\varphi] &= \mathcal{L}_{i\phi^*}(\phi, u, v_i)[\varphi] = \langle \bar{g}_i, \varphi \rangle \\ &= \int_{\Omega(\phi)} \bar{g}_{\zeta bi} \cdot \varphi \, dx + \int_{\partial\Omega(\phi)} \bar{g}_{\partial\Omega i} \cdot \varphi \, d\gamma + \int_{\Gamma_p(\phi)} \bar{g}_{pi} \cdot \varphi \, d\gamma \\ &\quad + \int_{\partial\Gamma_p(\phi) \cup \Theta_p(\phi)} \bar{g}_{\partial pi} \cdot \varphi \, d\zeta + \int_{\Gamma_{\eta i}(\phi)} \bar{g}_{\eta i} \cdot \varphi \, d\gamma \\ &\quad + \int_{\partial\Gamma_{\eta i}(\phi) \cup \Theta_{\eta i}(\phi)} \bar{g}_{\partial\eta i} \cdot \varphi \, d\zeta + \int_{\Gamma_D(\phi)} \bar{g}_{Di} \cdot \varphi \, d\gamma, \end{aligned}\quad (9.8.35)$$

where

$$\bar{g}_{\zeta bi} = \zeta_i \phi^* + u b^*, \quad (9.8.36)$$

$$\bar{g}_{\partial\Omega i} = (\zeta_i - \nabla u \cdot \nabla v_i + b v_i) \mathbf{v}, \quad (9.8.37)$$

$$\bar{g}_{pi} = \{\partial_v (p_N v_i) + \kappa p_N v_i\} \mathbf{v} + v_i p_N^*, \quad (9.8.38)$$

$$\bar{g}_{\partial pi} = p_N v_i \boldsymbol{\tau}, \quad (9.8.39)$$

$$\bar{g}_{\eta i} = (\partial_v \eta_{Ni} + \kappa \eta_{Ni}) \mathbf{v} + \eta_{Ni} \phi^*, \quad (9.8.40)$$

$$\bar{g}_{\partial\eta i} = \eta_{Ni} \boldsymbol{\tau}, \quad (9.8.41)$$

$$\bar{g}_{Di} = \{\partial_v (u - u_D) \partial_v v_i + \partial_v (v_i - v_{Di}) \partial_v u\} \mathbf{v} + \eta_{Di} \phi^*. \quad (9.8.42)$$

If \mathbf{g}_i of Eq. (9.8.5) and \bar{g}_i of Eq. (9.8.35) are compared in the case $\eta_{Di} \phi' = \mathbf{0}_{\mathbb{R}^d}$, although the term with \bar{g}_D of Eq. (9.8.42) appears on $\Gamma_D(\phi)$ in \bar{g}_i , there is no such component in \mathbf{g}_i . This result shows that if \mathbf{g}_i is used, and even when $\Gamma_D(\phi)$ varies, no additional treatment is needed.

Based on the results above, the following results can be obtained with respect to the function space containing \bar{g}_i of Eq. (9.8.35).

Theorem 9.8.6 (Shape Derivative \bar{g}_i of f_i) *Let $\phi \in \mathcal{D}$, b , p_N , u_D , ζ_i , η_{Ni} and η_{Di} be given functions fixed in space satisfying Hypotheses 9.5.2 and 9.6.2, and $\partial\Omega(\phi)$ be in the class of $H^3 \cap C^{1,1}$. Moreover, let u and v_i be the weak solutions of the state determination problem (Problem 9.5.4) and the adjoint problem (Problem 9.8.1) with respect to f_i , respectively, such that $u - u_D$ and $v_i - \eta_{Di} \partial_v u$ belong to $U(\phi) \cap W^{2,2q_{\mathbb{R}}}(D; \mathbb{R})$ ($q_{\mathbb{R}} > d$). When $\bar{g}_{\partial pi}$ and $\bar{g}_{\partial\eta i}$ in Eqs. (9.8.39) and (9.8.41), respectively, are zero, the shape derivative of f_i becomes \bar{g}_i in Eq. (9.8.35) and is*

an element of X' . Furthermore, we have

$$\begin{aligned}\bar{\mathbf{g}}_{\zeta bi}, \bar{\mathbf{g}}_{\partial\Omega i} &\in H^{1/2} \cap L^\infty \left(\partial\Omega(\boldsymbol{\phi}); \mathbb{R}^d \right), \\ \bar{\mathbf{g}}_{pi} &\in H^{1/2} \cap L^\infty \left(\Gamma_p(\boldsymbol{\phi}); \mathbb{R}^d \right), \\ \bar{\mathbf{g}}_{\eta i} &\in H^{1/2} \cap L^\infty \left(\Gamma_{\eta i}(\boldsymbol{\phi}); \mathbb{R}^d \right), \\ \bar{\mathbf{g}}_D &\in H^{1/2} \cap L^\infty \left(\Gamma_D(\boldsymbol{\phi}); \mathbb{R}^d \right).\end{aligned}\quad \square$$

Proof The fact that the shape derivative of f_i becomes Eq. (9.8.35) is as shown above. The regularity of $\bar{\mathbf{g}}_i$ can be shown by using relationships similar to the proof of Theorem 9.8.2 using the fact that u and v_i are in $W^{2,2qR}(D; \mathbb{R})$ and Hypothesis 9.6.2. \square

From the results of Theorem 9.8.2 and Theorem 9.8.6 the following can be said about the regularity of the shape optimization problem.

Remark 9.8.7 (Irregularity of Shape Optimization Problem) From Theorem 9.8.2 and Theorem 9.8.6, it was confirmed that \mathbf{g}_i and $\bar{\mathbf{g}}_i$ are both in X' with respect to X defined in Eq. (9.1.1). In other words, it is possible to define the Fréchet derivatives of cost functions with respect to the domain variation. However, it is not necessarily the case that \mathbf{g}_i and $\bar{\mathbf{g}}_i$ are in the linear space $H^2 \cap C^{0,1}(D; \mathbb{R}^d)$ containing the admissible set of design variables. This result indicates the fact that if $\boldsymbol{\varphi}$ is obtained by the gradient method substituting $-\mathbf{g}_i$ into $\boldsymbol{\varphi}$, $\boldsymbol{\varphi} + \boldsymbol{\varphi}$ is not guaranteed to be contained in $H^2 \cap C^{0,1}(D; \mathbb{R}^d)$, which is the linear space for the admissible set of design variables. This is thought to be a reason for the numerically unstable phenomena such as the rippling shapes explained at the start of this chapter. \square

9.9 Descent Directions of Cost Functions

Remark 9.8.7 points out the irregularity of the shape optimization problem of domain variation type. Hence, let us think about the gradient method and Newton method which both have the feature of regularizing the shape derivatives of cost functions in the framework of the abstract gradient and Newton methods on the linear space X of design variable. Here, let us assume that the gradient $\mathbf{g}_i \in X'$ of Eq. (9.8.5) and the Hessian $h_i \in \mathcal{L}^2(X \times X; \mathbb{R})$ of Eq. (9.8.24) with respect to the $i \in \{0, \dots, m\}$ th cost function f_i are given and think about the way to obtain the descent direction of f_i using the gradient method and Newton method on the linear space X of design variables.

9.9.1 H^1 Gradient Method

Choose a cost function $f_i(\phi, u)$ among $i \in \{0, \dots, m\}$ and assume that the shape derivative $g_i \in X'$ or $\tilde{g}_i \in X'$ at $\phi \in \mathcal{D}^\circ$ is given. From now on, $\tilde{f}_i(\phi) = f_i(\phi, u(\phi))$ will be denoted as $f_i(\phi)$. The method for obtaining the decent direction vector $\varphi_{g_i} \in X$ (domain variation) of f_i as the solution to the next problem is called an H^1 gradient method of domain variation type.

Problem 9.9.1 (H^1 Gradient Method of Domain Variation Type) Define X as in Eq. (9.1.1). Choose a coercive and bounded bilinear form $a_X : X \times X \rightarrow \mathbb{R}$ on X . In other words, suppose that there exist some positive constants α_X and β_X such that the inequalities

$$a_X(\varphi, \varphi) \geq \alpha_X \|\varphi\|_X^2, \quad |a_X(\varphi, \psi)| \leq \beta_X \|\varphi\|_X \|\psi\|_X \quad (9.9.1)$$

hold with respect to arbitrary $\varphi \in X$ and $\psi \in X$. Moreover, suppose $g_i \in X'$ is given at $\phi \in \mathcal{D}^\circ$. In this case, obtain $\varphi_{g_i} \in X$ which satisfies

$$a_X(\varphi_{g_i}, \psi) = -\langle g_i, \psi \rangle \quad (9.9.2)$$

for any $\psi \in X$. □

The way to choose $a_X : X \times X \rightarrow \mathbb{R}$ as in Problem 9.9.1 has arbitrary properties. Several specific examples will be shown in the section below.

Method Using the Inner Product in H^1 Space

Consider a method using the inner product on a real Hilbert space in a similar way to the H^1 gradient method of density variation type. In this case, it is allowed to assume that $\bar{\Omega}_{C0} = \emptyset$ on Eq. (9.1.1).

The inner product on $X = H^1(D; \mathbb{R}^d)$ is defined as

$$(\varphi, \psi)_X = \int_{\Omega(\phi)} \left\{ (\nabla \varphi^\top) \cdot (\nabla \psi^\top) + \varphi \cdot \psi \right\} dx$$

with respect to $\varphi \in X$ and $\psi \in X$. Let c_Ω be some positive-valued function contained in $L^\infty(D; \mathbb{R})$ such that

$$a_X(\varphi, \psi) = \int_{\Omega(\phi)} \left\{ (\nabla \varphi^\top) \cdot (\nabla \psi^\top) + c_\Omega \varphi \cdot \psi \right\} dx \quad (9.9.3)$$

is a bounded and coercive bilinear form on X . Here, c_Ω controls the weight of the first and second terms in the integrand. If c_Ω is taken to be small and the first term is made dominant, the smoothing function is prioritized. However, setting $c_\Omega = 0$

is not allowed, since the coercivity of the bilinear form will be lost, which is a requirement of the H^1 gradient method for it to hold. Moreover, if we write the symmetrical component of $\nabla\varphi^\top$ as

$$\mathbf{E}(\varphi) = (e_{ij}(\varphi))_{ij} = \frac{1}{2} \left\{ \nabla\varphi^\top + (\nabla\varphi^\top)^\top \right\},$$

the following bilinear form on X :

$$a_X(\varphi, \psi) = \int_{\Omega(\phi)} (\mathbf{E}(\varphi) \cdot \mathbf{E}(\psi) + c_\Omega \varphi \cdot \psi) \, dx \quad (9.9.4)$$

is also bounded and coercive. Excluding antisymmetric components of $\nabla\varphi^\top$ indicates rotational motion which does not generate deformation.

Furthermore, $\mathbf{C} = (c_{ijkl})_{ijkl} \in L^\infty(D; \mathbb{R}^{d \times d \times d \times d})$ is taken to be a stiffness tensor used in a linear elastic problem. In other words, we assume that there exist positive constants α_X and β_X such that the bounds

$$\mathbf{A} \cdot (\mathbf{C}\mathbf{A}) \geq \alpha_X \|\mathbf{A}\|^2, \quad |\mathbf{A} \cdot (\mathbf{C}\mathbf{B})| \leq \beta_X \|\mathbf{A}\| \|\mathbf{B}\| \quad (9.9.5)$$

hold for any symmetric tensors $\mathbf{A} \in \mathbb{R}^{d \times d}$ and $\mathbf{B} \in \mathbb{R}^{d \times d}$ and that the symmetry condition $c_{ijkl} = c_{klji}$ also holds. Using these, let the stress tensor be

$$\mathbf{S}(\varphi) = \mathbf{C}\mathbf{E}(\varphi) = \left(\sum_{(k,l) \in \{1, \dots, d\}^2} c_{ijkl} e_{kl}(\varphi) \right)_{ij}. \quad (9.9.6)$$

In this case, we have that

$$a_X(\varphi, \psi) = \int_{\Omega(\phi)} (\mathbf{S}(\varphi) \cdot \mathbf{E}(\psi) + c_\Omega \varphi \cdot \psi) \, dx \quad (9.9.7)$$

is a bounded and coercive bilinear form on X . $a_X(\varphi, \psi)$ of Eq. (9.9.7) is a bilinear form providing the variation of strain energy in a linear elastic problem when φ and ψ are viewed as the displacement and its variation. In this case, c_Ω indicates the spring constant of the distributed spring placed in D . Figure 9.13 provides an illustration of Problem 9.9.1 in this case.

Figure 9.13a represents the case when \mathbf{g}_i of Eq. (9.8.5) is used as the shape derivative of f_i . Problem 9.9.1 in this case is given by a weak-form equation. Hence, if we dare to rewrite this problem in its strong form, the following assumptions are needed. When u and v_i are elements of $W^{2,2q_R}$, the first term on the right-hand side

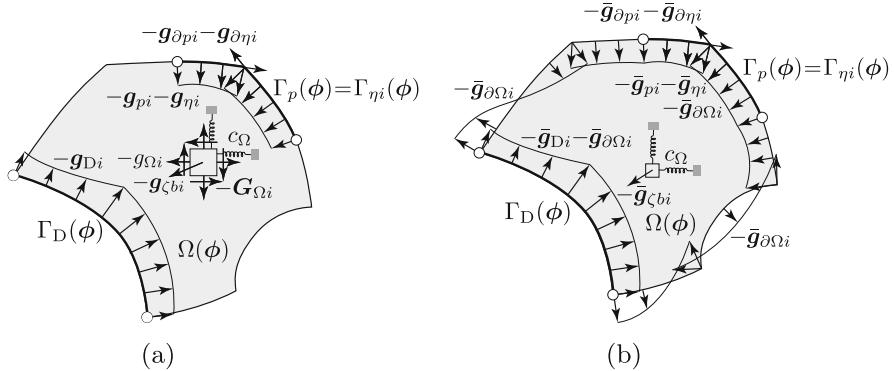


Fig. 9.13 The H^1 gradient method using an inner product of $H^1(D; \mathbb{R}^d)$. **(a)** When using \mathbf{g}_i . **(b)** When using $\tilde{\mathbf{g}}_i$

of Eq. (9.8.5) is written as

$$\begin{aligned}
& \int_{\Omega(\phi)} \left\{ \mathbf{G}_{\Omega i} \cdot (\nabla \boldsymbol{\varphi}^\top) + g_{\Omega i} \nabla \cdot \boldsymbol{\varphi} + \mathbf{g}_{\zeta bi} \cdot \boldsymbol{\varphi} \right\} dx \\
&= \int_{\Omega(\phi)} \left\{ \nabla \cdot (\mathbf{G}_{\Omega i} \boldsymbol{\varphi}) - (\nabla^\top \mathbf{G}_{\Omega i})^\top \cdot \boldsymbol{\varphi} + \nabla \cdot (g_{\Omega i} \boldsymbol{\varphi}) \right. \\
&\quad \left. - (\nabla g_{\Omega i}) \cdot \boldsymbol{\varphi} + \mathbf{g}_{\zeta bi} \cdot \boldsymbol{\varphi} \right\} dx \\
&= \int_{\Omega(\phi)} \tilde{\mathbf{g}}_{\Omega i} \cdot \boldsymbol{\varphi} dx + \int_{\partial\Omega(\phi)} \tilde{\mathbf{g}}_{\partial\Omega i} \cdot \boldsymbol{\varphi} d\gamma, \tag{9.9.8}
\end{aligned}$$

where

$$\tilde{\mathbf{g}}_{\Omega i} = - \left(\nabla^\top \mathbf{G}_{\Omega i} \right)^\top - \nabla g_{\Omega i} + \mathbf{g}_{\zeta bi}, \quad (9.9.9)$$

$$\tilde{\mathbf{g}}_{\partial\Omega i} = (\mathbf{G}_{\Omega i} + g_{\Omega i}) \mathbf{v}. \quad (9.9.10)$$

Moreover, $\chi_{\Gamma_p(\phi)} : \partial\Omega(\phi) \rightarrow \mathbb{R}$ represents the characteristic function which takes the value 1 on $\Gamma_p(\phi) \subset \partial\Omega(\phi)$ and value 0 on $\partial\Omega(\phi) \setminus \bar{\Gamma}_p(\phi)$. In this case, the strong form of Problem 9.9.1 using Eq. (9.9.7) for $a_X(\varphi, \psi)$ is given as follows.

Problem 9.9.2 (H^1 Gradient Method Using H^1 Inner Product and g_i) Let g_{pi} , $g_{\partial pi}$, $g_{\eta i}$, $g_{\partial \eta i}$ and g_{Di} of Eq. (9.8.5) as well as $\tilde{g}_{\Omega i}$ and $\tilde{g}_{\partial \Omega i}$ of Eqs. (9.9.9) and

(9.9.10), respectively, be given at $\phi \in \mathcal{D}^\circ$. Find $\varphi_{gi} : \Omega(\phi) \rightarrow \mathbb{R}$ which satisfies

$$-\nabla^\top S(\varphi_{gi}) + c_\Omega \varphi_{gi}^\top = -\tilde{\mathbf{g}}_{\Omega i}^\top \quad \text{in } \Omega(\phi), \quad (9.9.11)$$

$$\begin{aligned} S(\varphi_{gi}) \mathbf{v} &= -\chi_{\Gamma_p(\phi)} \mathbf{g}_{pi} - \chi_{\Gamma_{\eta i}(\phi)} \mathbf{g}_{\eta i} - \chi_{\Gamma_D(\phi)} \mathbf{g}_{Di} - \tilde{\mathbf{g}}_{\partial \Omega i} \\ &\quad \text{on } \partial \Omega(\phi), \end{aligned} \quad (9.9.12)$$

$$\begin{aligned} S(\varphi_{gi}) \boldsymbol{\tau} &= -\chi_{\partial \Gamma_p(\phi) \cup \Theta_p(\phi)} \mathbf{g}_{\partial pi} - \chi_{\partial \Gamma_{\eta i}(\phi) \cup \Theta_{\eta i}(\phi)} \mathbf{g}_{\partial \eta i} \\ &\quad \text{on } \partial \Omega(\phi). \end{aligned} \quad (9.9.13)$$

□

Figure 9.13b shows the case when the shape gradient $\tilde{\mathbf{g}}_i$ of f_i is given by Eq. (9.8.35). The strong form in this case is given as follows.

Problem 9.9.3 (H^1 Gradient Method Using H^1 Inner Product and $\tilde{\mathbf{g}}_i$) Let $\tilde{\mathbf{g}}_{\zeta bi}$, $\tilde{\mathbf{g}}_{\partial \Omega i}$, $\tilde{\mathbf{g}}_{pi}$, $\tilde{\mathbf{g}}_{\partial pi}$, $\tilde{\mathbf{g}}_{\eta i}$, $\tilde{\mathbf{g}}_{\partial \eta i}$ and $\tilde{\mathbf{g}}_{Di}$ as in Eq. (9.8.35) be given at $\phi \in \mathcal{D}^\circ$, obtain φ_{gi} which satisfies

$$-\nabla^\top S(\varphi_{gi}) + c_\Omega \varphi_{gi}^\top = -\tilde{\mathbf{g}}_{\zeta bi}^\top \quad \text{in } \Omega(\phi),$$

$$\begin{aligned} S(\varphi_{gi}) \mathbf{v} &= -\chi_{\Gamma_p(\phi)} \tilde{\mathbf{g}}_{pi} - \chi_{\Gamma_{\eta i}(\phi)} \tilde{\mathbf{g}}_{\eta i} - \chi_{\Gamma_D(\phi)} \tilde{\mathbf{g}}_{Di} - \tilde{\mathbf{g}}_{\partial \Omega i} \\ &\quad \text{on } \partial \Omega(\phi), \end{aligned}$$

$$\begin{aligned} S(\varphi_{gi}) \boldsymbol{\tau} &= -\chi_{\partial \Gamma_p(\phi) \cup \Theta_p(\phi)} \tilde{\mathbf{g}}_{\partial pi} - \chi_{\partial \Gamma_{\eta i}(\phi) \cup \Theta_{\eta i}(\phi)} \tilde{\mathbf{g}}_{\partial \eta i} \\ &\quad \text{on } \partial \Omega(\phi). \end{aligned}$$

□

Method Using Boundary Condition

Moreover, as with the H^1 gradient method of varying density type, the bilinear form $a_X : X \times X \rightarrow \mathbb{R}$ can be made coercive by adding a boundary condition.

Firstly, think about using a Dirichlet boundary condition. In Eq. (9.1.1), $\bar{\Omega}_{C0} \subset \bar{\Omega}_0$ was defined as a boundary or a closure of domain on which the domain variation is fixed as a design demand. Here, $|\bar{\Omega}_{C0}| > 0$ is assumed. In this case,

$$a_X(\varphi, \psi) = \int_{\Omega(\phi) \setminus \bar{\Omega}_{C0}} S(\varphi) \cdot E(\psi) \, dx \quad (9.9.14)$$

is a bounded and coercive bilinear form on X . This is because, when the measure of $\bar{\Omega}_{C0}$ is positive and $\varphi = \mathbf{0}_{\mathbb{R}^d}$ on $\bar{\Omega}_{C0}$, Korn's inequality implies that there exists a positive constant c which depends only on $\Omega(\phi) \setminus \bar{\Omega}_{C0}$ such that the inequality

$$a_X(\varphi, \varphi) \geq \alpha_X \|E(\varphi)\|_{L^2(\Omega(\phi) \setminus \bar{\Omega}_{C0}; \mathbb{R}^{d \times d})}^2 \geq c \|\varphi\|_{H^1(\Omega(\phi) \setminus \bar{\Omega}_{C0}; \mathbb{R}^d)}^2$$

holds. Here α_X is a positive constant satisfying Eq. (9.9.5). The strong form in this case is shown as follows. Here, the situation when the shape gradient of f_i is given by $\bar{\mathbf{g}}_i$ in Eq. (9.8.35) is shown.

Problem 9.9.4 (H^1 Gradient Method Using Dirichlet Condition and $\bar{\mathbf{g}}_i$) Let $\bar{\mathbf{g}}_{\zeta bi}$, $\bar{\mathbf{g}}_{\partial\Omega i}$, $\bar{\mathbf{g}}_{pi}$, $\bar{\mathbf{g}}_{\partial pi}$, $\bar{\mathbf{g}}_{\eta i}$, $\bar{\mathbf{g}}_{\partial\eta i}$ and $\bar{\mathbf{g}}_{Di}$ as in Eq. (9.8.35) be given at $\phi \in \mathcal{D}^0$. Obtain $\varphi_{gi} : \Omega(\phi) \setminus \bar{\Omega}_{C0} \rightarrow \mathbb{R}^d$ which satisfies

$$\begin{aligned} -\nabla^\top S(\varphi_{gi}) + c_\Omega \varphi_{gi}^\top &= -\bar{\mathbf{g}}_{\zeta bi}^\top \quad \text{in } \Omega(\phi) \setminus \bar{\Omega}_{C0}, \\ S(\varphi_{gi}) \mathbf{v} &= -\chi_{\Gamma_p(\phi)} \bar{\mathbf{g}}_{pi} - \chi_{\Gamma_{\eta i}(\phi)} \bar{\mathbf{g}}_{\eta i} - \chi_{\Gamma_D(\phi)} \bar{\mathbf{g}}_{Di} - \bar{\mathbf{g}}_{\partial\Omega i} \\ &\quad \text{on } \partial\Omega(\phi) \setminus \bar{\Omega}_{C0}, \\ S(\varphi_{gi}) \boldsymbol{\tau} &= -\chi_{\partial\Gamma_p(\phi) \cup \Theta_p(\phi)} \bar{\mathbf{g}}_{\partial pi} - \chi_{\partial\Gamma_{\eta i}(\phi) \cup \Theta_{\eta i}(\phi)} \bar{\mathbf{g}}_{\partial\eta i} \\ &\quad \text{on } \partial\Omega(\phi) \setminus \bar{\Omega}_{C0}, \\ \varphi_{gi} &= \mathbf{0}_{\mathbb{R}^d} \quad \text{on } \bar{\Omega}_{C0}. \end{aligned}$$

□

Figure 9.14a illustrates Problem 9.9.4. This problem assumes that $\Omega(\phi)$ is a linear elastic body and obtains the displacement φ_{gi} when $\bar{\Omega}_{C0}$ is fixed, the remaining boundaries are applied with the traction containing $-\bar{\mathbf{g}}_{\partial\Omega i}$, $-\bar{\mathbf{g}}_{pi}$, $-\bar{\mathbf{g}}_{\partial pi}$, $-\bar{\mathbf{g}}_{\eta i}$, $-\bar{\mathbf{g}}_{\partial\eta i}$ and $-\bar{\mathbf{g}}_{Di}$, and the volume force $-\bar{\mathbf{g}}_{\zeta bi}$ is applied. From this sort of interpretation, Problem 9.9.4 is known as the traction method[8].

Furthermore, if the Robin condition is used, even if $\bar{\Omega}_{C0} = \emptyset$ is assumed in Eq. (9.1.1), coerciveness of $a_X(\varphi, \psi)$ can be obtained. Choose some positive-valued function $c_{\partial\Omega} \in L^\infty(\partial\Omega(\phi); \mathbb{R})$ and let

$$a_X(\varphi, \psi) = \int_{\Omega(\phi)} S(\varphi) \cdot E(\psi) \, dx + \int_{\partial\Omega(\phi)} c_{\partial\Omega}(\varphi \cdot \mathbf{v}) (\psi \cdot \mathbf{v}) \, d\gamma. \quad (9.9.15)$$

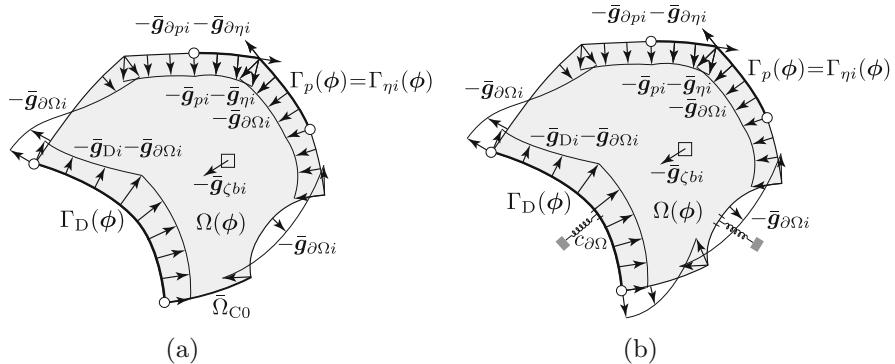


Fig. 9.14 The H^1 gradient method (when $\Gamma_p(\phi) = \Gamma_{\eta i}(\phi)$). (a) Dirichlet condition. (b) Robin condition

The strong form in this case is shown below. Here too, let us show only the case when the shape gradient of f_i is given by $\bar{\mathbf{g}}_i$ in Eq. (9.8.35).

Problem 9.9.5 (H^1 Gradient Method Using Robin Condition and $\bar{\mathbf{g}}_i$) Let $\bar{\mathbf{g}}_{\zeta bi}$, $\bar{\mathbf{g}}_{\partial\Omega i}$, $\bar{\mathbf{g}}_{pi}$, $\bar{\mathbf{g}}_{\partial pi}$, $\bar{\mathbf{g}}_{\eta i}$, $\bar{\mathbf{g}}_{\partial\eta i}$ and $\bar{\mathbf{g}}_{Di}$ as in Eq. (9.8.35) be given at $\phi \in \mathcal{D}^\circ$. Find φ_{gi} which satisfies

$$\begin{aligned} -\nabla^\top \mathbf{S}(\varphi_{gi}) &= -\bar{\mathbf{g}}_{\zeta bi}^\top \quad \text{in } \Omega(\phi), \\ \mathbf{S}(\varphi_{gi}) \mathbf{v} + c_{\partial\Omega}(\varphi \cdot \mathbf{v}) \mathbf{v} &= -\chi_{\Gamma_p(\phi)} \bar{\mathbf{g}}_{pi} - \chi_{\Gamma_{\eta i}(\phi)} \bar{\mathbf{g}}_{\eta i} \\ &\quad - \chi_{\Gamma_D(\phi)} \bar{\mathbf{g}}_{Di} - \bar{\mathbf{g}}_{\partial\Omega i} \quad \text{on } \partial\Omega(\phi), \\ \mathbf{S}(\varphi_{gi}) \boldsymbol{\tau} &= -\chi_{\partial\Gamma_p(\phi) \cup \Theta_p(\phi)} \bar{\mathbf{g}}_{\partial pi} - \chi_{\partial\Gamma_{\eta i}(\phi) \cup \Theta_{\eta i}(\phi)} \bar{\mathbf{g}}_{\partial\eta i} \\ &\quad \text{on } \partial\Omega(\phi). \end{aligned}$$

□

Figure 9.14b illustrates Problem 9.9.5. This problem assumes that $\Omega(\phi)$ is a linear elastic body and that a distribution spring with spring constant $c_{\partial\Omega}$ is placed on $\partial\Omega(\phi)$, then seeks the displacement φ_{gi} when $\bar{\Omega}_{C0}$ is fixed, the remaining boundaries are applied with the traction containing $-\bar{\mathbf{g}}_{\partial\Omega i}$, $-\bar{\mathbf{g}}_{pi}$, $-\bar{\mathbf{g}}_{\partial pi}$, $-\bar{\mathbf{g}}_{\eta i}$, $-\bar{\mathbf{g}}_{\partial\eta i}$ and $-\bar{\mathbf{g}}_{Di}$, and the volume force $-\bar{\mathbf{g}}_{\zeta bi}$ is applied. From this sort of interpretation, Problem 9.9.5 has been referred to as the traction method with spring, or traction method of Robin type[14].

Regularity of the H^1 Gradient Method

From the weak solutions of Problem 9.9.1 and its specific examples (Problems 9.9.3 to 9.9.5), the results below can be obtained. Here, we call the neighborhood of points or edges such that u does not belong to \mathcal{S} (or $U(\phi) \cap W^{2,2qR}(D; \mathbb{R})$ when $\bar{\mathbf{g}}_i$ in Theorem 9.8.6 is used) a neighborhood of singular points and will write it as $B(\phi)$ (refer to Hypothesis 9.5.3). Also, we let $f_i(\phi, u)$ be denoted by $\tilde{f}_i(\phi)$ when u is the solution to Problem 9.5.4.

Theorem 9.9.6 (Regularity of the H^1 Gradient Method) *There exists a unique weak solution $\varphi_{gi} \in X$ for each Problem 9.9.2 to 9.9.5 using \mathbf{g}_i of Theorem 9.8.2 or $\bar{\mathbf{g}}_i$ of Theorem 9.8.6. φ_{gi} is a function of class $H^2 \cap C^{0,1}$ on $\Omega(\phi) \setminus \bar{B}(\phi)$. Moreover, φ_{gi} points to the direction of the domain variation which decreases the value of $\tilde{f}_i(\phi)$.* □

Proof Let us think about the weak solution φ_{gi} of Problem 9.9.2. Problem 9.9.2 is a boundary value problem of an elliptic partial differential equation with $\mathbf{G}_{\Omega i}$ and $\mathbf{g}_{\Omega i}$ of class $H^1 \cap L^\infty$ in Theorem 9.8.2 given in the domain, and \mathbf{g}_{pi} , $\mathbf{g}_{\partial pi}$, $\mathbf{g}_{\eta i}$ and $\mathbf{g}_{\partial\eta i}$ of class $H^{1/2} \cap L^\infty$ in Theorem 9.8.2 given as Neumann boundary conditions. Hence, $\varphi_{gi} \in X$ exists uniquely from the Lax-Milgram theorem. Moreover, φ_{gi} is of class $H^2 \cap C^{0,1}$ on $\Omega(\phi) \setminus \bar{B}(\phi)$. This is due to the fact that $\mathbf{G}_{\Omega i}$ and $\mathbf{S}(\varphi_{gi})$ in

Eq. (9.9.11) have the same regularity, and thus $\mathbf{G}_{\Omega i}$ being of class $H^1 \cap L^\infty$ means φ_{gi} is of class $H^2 \cap C^{0,1}$. In fact, using Eqs. (9.9.12) and (9.9.13), one can also infer that φ_{gi} is of class $H^2 \cap C^{0,1}$.

Similarly, the weak solution φ_{gi} of Problem 9.9.3 satisfies an elliptic partial differential equation with $\bar{\mathbf{g}}_{\partial\Omega i}$, $\bar{\mathbf{g}}_{pi}$, $\bar{\mathbf{g}}_{\partial pi}$, $\bar{\mathbf{g}}_{\eta i}$, $\bar{\mathbf{g}}_{\partial\eta i}$ and $\bar{\mathbf{g}}_{Di}$ of class $H^{1/2} \cap L^\infty$ in Theorem 9.8.6 as Neumann boundary conditions. Here, there exists a unique weak solution $\varphi_{gi} \in X$ from the Lax–Milgram theorem being of class $H^2 \cap C^{0,1}$ on $\Omega(\phi) \setminus \bar{B}(\phi)$. Similar results can be obtained for the weak solutions of Problem 9.9.4 and Problem 9.9.5.

Furthermore, with respect to the weak solutions φ_{gi} of Problems 9.9.3 to 9.9.5, we have the estimate

$$\begin{aligned}\tilde{f}_i(\phi + \bar{\epsilon}\varphi_{gi}) - \tilde{f}_i(\phi) &= \bar{\epsilon}\langle \mathbf{g}_i, \varphi_{gi} \rangle + o(|\bar{\epsilon}|) \\ &= -\bar{\epsilon}\alpha_X(\varphi_{gi}, \varphi_{gi}) + o(|\bar{\epsilon}|) \leq -\bar{\epsilon}\alpha_X \|\varphi_{gi}\|_X^2 + o(|\bar{\epsilon}|)\end{aligned}$$

for some positive constant $\bar{\epsilon}$. Hence, if $\|\varphi_{gi}\|_X$ is taken to be sufficiently small, $\tilde{f}_i(\phi)$ is reduced. \square

The following remark can be made about the relationship between the result of Theorem 9.9.6 and the admissible set \mathcal{D} of domain variations defined by Eq. (9.1.3).

Remark 9.9.7 (H^1 Gradient Method for Shape Optimization Problem) From Theorem 9.9.6, it was confirmed that the domain variation φ_{gi} obtained by the H^1 gradient method with respect to the shape optimization problem is contained in the linear space $H^2 \cap C^{0,1}(D; \mathbb{R}^d)$ for the admissible set \mathcal{D} of design variables excepting the neighborhood of singular points. From this, the domain can be moved via continuous mapping excluding the neighborhood of singular points. However, one cannot guarantee to have the bound $|\phi + \varphi_{gi}|_{C^{0,1}(D; \mathbb{R}^d)} \leq \sigma$ or the condition that $\tilde{\Gamma}(\phi + \varphi_{gi})$ ($\tilde{\Gamma}_0 = \Gamma_{p0} \cup \Gamma_{\eta00} \cup \Gamma_{\eta10} \cup \dots \cup \Gamma_{\eta m0} \setminus \bar{\Omega}_{C0}$) is of class $H^3 \cap C^{1,1}$ which are sufficient conditions for the inverse mapping of $\phi + \varphi_{gi}$ to become bijective. If a numerically unstable phenomenon caused by these conditions not being satisfied occurs, there is a need to consider additional requirements in order to satisfy the said conditions. \square

As one of methods to improve the regularity of the boundary, an iterative method of the H^1 gradient method can be considered. This method is the following algorithm.

Algorithm 9.1 (Iterative Method of the H^1 Gradient Method) Let a domain $\Omega(\phi)$ be given. Obtain a domain variation in the following way:

- (1) Calculate a shape gradient \mathbf{g}_i (or $\bar{\mathbf{g}}_i$).
- (2) By the first H^1 gradient method, obtain $\varphi_{gi} = \varphi_{gi1}$ using $-\mathbf{g}_i$ (or $-\bar{\mathbf{g}}_i$). Here, φ_{gi1} is not used for domain variation.

- (3) Using the trace $\varphi_{gi1}|_{\partial\Omega(\phi)}$ of the solution φ_{gi1} of the first H^1 gradient method on $\partial\Omega(\phi)$ instead of $-\bar{g}_i$, calculate φ_{gi2} by the second H^1 gradient method. Vary the domain $\Omega(\phi)$ with the φ_{gi2} .

For the boundary of the new domain $\Omega(\phi + \varphi_{gi2})$ obtained in the above way, it is expected that the differentiability improves by one order higher than $\Omega(\phi + \varphi_{gi1})$.

9.9.2 H^1 Newton Method

Now, if the second-order derivative (Hessian) $h_i \in \mathcal{L}^2(X \times X; \mathbb{R})$ of the cost function f_i is computable, a Newton method on $X = H^1(D; \mathbb{R}^d)$ can be considered. This method is called an H^1 Newton method of domain variation type.

Problem 9.9.8 (H^1 Newton Method of Domain Variation Type) Let X and \mathcal{D} be given by Eqs. (9.1.1) and (9.1.3), respectively. Let the shape derivative and second-order shape derivative of $f_i \in C^2(\mathcal{D}; \mathbb{R})$ at $\phi_k \in \mathcal{D}^\circ$ which is not a local minimizer be $g_i(\phi_k) \in X'$ and $h_i(\phi_k) \in \mathcal{L}^2(X \times X; \mathbb{R})$, respectively. Moreover, assume that $a_X : X \times X \rightarrow \mathbb{R}$ is a bilinear form which assures coercivity and sufficient regularity of $h_i(\phi_k)$ on X . In this case, obtain $\varphi_{gi} \in X$ which satisfies

$$h_i(\phi_k)[\varphi_{gi}, \psi] + a_X(\varphi_{gi}, \psi) = -\langle g_i(\phi_k), \psi \rangle \quad (9.9.16)$$

with respect to an arbitrary $\psi \in X$. □

In Problem 9.9.8, if the Newton method is considered with only the expression for h_i appearing on the left-hand side of Eq. (9.9.16), there may be cases when the coerciveness of h_i on X may not be guaranteed. In reality, h_i calculated by Eq. (9.8.24) contains a negative term, hence, the addition of the bilinear form a_X which is bounded and coercive on X to the left-hand side of Eq. (9.9.16) in Problem 9.9.8. For instance, in the case using the inner product on X such as Eq. (9.9.3), we can assume

$$a_X(\varphi, \psi) = \int_{\Omega(\phi)} \left\{ c_{\Omega 1} (\nabla \varphi^\top) \cdot (\nabla \psi^\top) + c_{\Omega 0} \varphi \cdot \psi \right\} dx. \quad (9.9.17)$$

Here, $c_{\Omega 0}$ and $c_{\Omega 1}$ are positive constants for achieving coercivity for a_X and desired regularity for φ_{gi} in Eq. (9.9.16), respectively. $c_{\Omega 0}$ has the same meaning as that explained after Eq. (8.6.3) in Chap. 8.

Furthermore, in the case of the Newton method when the second-order shape derivative of $f_i(\phi)$ is given by the Hesse gradient, Problem 9.9.8 is replaced with the following problem.

Problem 9.9.9 (Newton Method Using Hesse Gradient) Under the assumption of Problem 9.9.8, the gradient of the shape derivative of f_i , a search vector and

the Hesse gradient of f_i at a non-local minimum point $\phi_k \in \mathcal{D}^\circ$ are denoted by $\mathbf{g}_i(\phi_k) \in X'$, $\bar{\phi}_{gi} \in X$ and $\mathbf{g}_{Hi}(\phi_k, \bar{\phi}_{gi}) \in X'$, respectively. Given a coercive and bounded bilinear form $a_X : X \times X \rightarrow \mathbb{R}$ on X , find a $\varphi_{gi} \in X$ which satisfies

$$a_X(\varphi_{gi}, \psi) = -\langle (\mathbf{g}_i(\phi_k) + \mathbf{g}_{Hi}(\phi_k, \bar{\phi}_{gi})), \psi \rangle \quad (9.9.18)$$

with respect to an arbitrary $\psi \in X$. \square

9.10 Solution to Shape Optimization Problem of Domain Variation Type

The shape optimization problem (Problem 9.6.3) of domain variation type has a correspondence with the abstract optimal design problem, as shown in Table 9.1. Therefore, the gradient method with respect to constrained problems shown in Sect. 7.7.1 (Sect. 3.7) and the Newton method with respect to a constrained problem shown in Sect. 7.7.2 (Sect. 3.8) are applicable as similarly shown in Chap. 8.

9.10.1 Gradient Method for Constrained Problems

The gradient method with respect to constrained problems employs a simple numerical procedure such as that given in Algorithm 3.6 shown in Sect. 3.7.1 with only a few modifications as follows:

- (1) The design variable x and its variation y are replaced by ϕ and φ , respectively.
- (2) The equation (Eq. (3.7.10)) that describes the gradient method is replaced with a condition such that there holds the equation

$$c_a a_X(\varphi_{gi}, \psi) = -\langle \mathbf{g}_i, \psi \rangle \quad (9.10.1)$$

for any $\psi \in X$, where $a_X(\varphi_{gi}, \psi)$ is a bilinear form on X used in the weak form of one of Problems 9.9.2 to 9.9.5.

Table 9.1 Correspondence between abstract optimal design problem and shape optimization problem of domain variation type

	Abstract problem	Domain variation type problem
Design variable	$\phi \in X$	$\phi \in X = H^1(D; \mathbb{R}^d)$
State variable	$u \in U$	$u \in U = H^1(D; \mathbb{R})$
Fréchet derivative of f_i	$g_i \in X'$	$\mathbf{g}_i \in X' = H^{1'}(D; \mathbb{R}^d)$
Solution of gradient method	$\varphi_{gi} \in X$	$\varphi_{gi} \in X = H^1(D; \mathbb{R}^d)$

(3) The equation (Eq. (3.7.11)) used to seek for the search vector is replaced with

$$\boldsymbol{\varphi}_g = \boldsymbol{\varphi}_{g0} + \sum_{i \in I_A} \lambda_i \boldsymbol{\varphi}_{gi}. \quad (9.10.2)$$

(4) The equation (Eq. (3.7.12)) used to seek for the Lagrange multiplier is replaced with

$$(\langle \mathbf{g}_i, \boldsymbol{\varphi}_{gj} \rangle)_{(i,j) \in I_A^2} (\lambda_j)_{j \in I_A} = - (f_i + \langle \mathbf{g}_i, \boldsymbol{\varphi}_{g0} \rangle)_{i \in I_A}. \quad (9.10.3)$$

Furthermore, if instead a complicated numerical procedure such as that given by Algorithm 3.7 is used, the following changes are added in addition to (1) to (4) above:

(5) Replace the Armijo criteria Eq. (3.7.26) with

$$\mathcal{L}(\boldsymbol{\phi} + \boldsymbol{\varphi}_g, \boldsymbol{\lambda}_{k+1}) - \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\lambda}) \leq \xi \left\langle \mathbf{g}_0 + \sum_{i \in I_A} \lambda_i \mathbf{g}_i, \boldsymbol{\varphi}_g \right\rangle, \quad (9.10.4)$$

where $\xi \in (0, 1)$.

(6) Replace the Wolfe criteria Eq. (3.7.27) with

$$\begin{aligned} & \mu \left\langle \mathbf{g}_0 + \sum_{i \in I_A} \lambda_i \mathbf{g}_i, \boldsymbol{\varphi}_g \right\rangle \\ & \leq \left\langle \mathbf{g}_0 (\boldsymbol{\phi} + \boldsymbol{\varphi}_g) + \sum_{i \in I_A} \lambda_{i+1} \mathbf{g}_i (\boldsymbol{\phi} + \boldsymbol{\varphi}_g), \boldsymbol{\varphi}_g \right\rangle, \end{aligned} \quad (9.10.5)$$

where μ is such that $0 < \xi < \mu < 1$.

(7) Replace the update equation for $\boldsymbol{\lambda}$ from the Newton–Raphson method given in Eq. (3.7.21) with

$$(\delta \lambda_j)_{j \in I_A} = - (\langle \mathbf{g}_i(\boldsymbol{\lambda}), \boldsymbol{\varphi}_{gj} \rangle)_{(i,j) \in I_A^2}^{-1} (f_i(\boldsymbol{\lambda}))_{i \in I_A}. \quad (9.10.6)$$

9.10.2 Newton Method for Constrained Problems

If, in addition to the first-order shape derivative, the corresponding second-order shape derivative of a cost function is also computable, the gradient method can be improved to the Newton method to numerically solve the associated constrained problem. In this case, we substitute $h_i(\boldsymbol{\phi}_k)[\boldsymbol{\varphi}_{gi}, \boldsymbol{\psi}]$ in Eq. (9.9.16) with the Hessian of the Lagrange function \mathcal{L} with respect to the shape optimization problem

(Problem 9.6.3) with

$$h_{\mathcal{L}}(\phi_k)[\varphi_{gi}, \psi] = h_0(\phi_k)[\varphi_{gi}, \psi] + \sum_{i \in I_A(\phi_k)} \lambda_{ik} h_i(\phi_k)[\varphi_{gi}, \psi]. \quad (9.10.7)$$

In other words, we let Eq. (9.9.16) be replaced with

$$c_h h_{\mathcal{L}}(\phi_k)[\varphi_{gi}, \psi] + a_X(\varphi_{gi}, \psi) = -\langle g_i(\phi_k), \psi \rangle, \quad (9.10.8)$$

where c_h and c_a are constants to control the step size. In this case, the simple Algorithm 3.8 shown in Sect. 3.8.1 can be used by applying the following substitution:

- (1) Replace the design variable x and its variation y by ϕ and φ , respectively.
- (2) Replace Eq. (3.7.10) with the solution of Eq. (9.10.8).
- (3) Replace Eq. (3.7.11) with Eq. (9.10.2).
- (4) Replace Eq. (3.7.12) with Eq. (9.10.3).

When the second-order shape derivative of $f_i(\phi)$ is obtained as a Hesse gradient, Eqs. (9.10.7) and (9.10.8) are replaced with

$$g_{H\mathcal{L}}(\phi_k, \bar{\varphi}_g) = g_{H0}(\phi_k, \bar{\varphi}_g) + \sum_{i \in I_A(\phi_k)} \lambda_{ik} g_{Hi}(\phi_k, \bar{\varphi}_g) \quad (9.10.9)$$

$$a_X(\varphi_{gi}, \psi) = -\langle (g_i(\phi_k) + c_h g_{H\mathcal{L}}(\phi_k, \bar{\varphi}_g)), \psi \rangle, \quad (9.10.10)$$

respectively. Using the definitions, the following step is added:

- (5) Replace Eq. (3.8.11) with Eq. (9.10.10).

If instead one wishes to implement a more complicated numerical procedure such as that shown in Sect. 3.8.2, then several additional requirements are needed in response to the added functionality and characteristics of such problems as those examined in Chap. 8.

9.11 Error Estimation

When the shape optimization problem (Problem 9.6.3) of domain variation type is to be solved using an algorithm such as that shown in Sect. 9.10, the search vector φ_g can be obtained by Eq. (9.10.2). For this purpose, there is a need to seek the numerical solutions of u for the state determination problem (Problem 9.5.4), the numerical solutions of $v_0, v_{i_1}, \dots, v_{i_{|I_A|}}$ for the adjoint problems with respect to $f_0, f_{i_1}, \dots, f_{i_{|I_A|}}$ (Problem 9.8.1), as well as the numerical solutions of $\varphi_0, \varphi_{i_1}, \dots, \varphi_{i_{|I_A|}}$ for the H^1 gradient method (Problem 9.9.1). The Lagrange multipliers $\lambda_{i_1},$

$\dots, \lambda_{i|I_A|}$ are calculated using these numerical solutions. As in Chap. 8, we assume here too that a finite element method is used to obtain the numerical solutions for the three types of boundary value problems. We then use the estimated error from the numerical solutions via the finite element method seen in Sect. 6.6 in order to conduct an error estimation for the search vector φ_g [120, 122].

In the case of the shape optimization problem of domain variation type, the defined domain of the boundary value problem is perturbed. Here, a situation is considered in which $\Omega(\phi)$ is assumed to be given and $\Omega(\phi + \varphi)$ is sought. In this section, for simplicity, we write $\Omega(\phi)$ as Ω . Similarly, $(\cdot)(\phi)$ is denoted by (\cdot) . Assume that Ω is a polyhedron (Sect. 6.6.1) and consider a regular finite element division $\mathcal{T} = \{\Omega_i\}_{i \in \mathcal{E}}$ with respect to Ω . Moreover, define the diameter h of the finite element as $h(\mathcal{T})$ of Eq. (6.6.2) and consider the finite element division sequence $\{\mathcal{T}_h\}_{h \rightarrow 0}$. The notations we give below will be used in the rest of the discussion:

- (1) The exact solution of the state determination problem (Problem 9.5.4) and the adjoint problems with respect to f_i (Problem 9.8.1) are written as u and $v_0, v_{i_1}, \dots, v_{i|I_A|}$, respectively. These numerical solutions from the finite element method are written as

$$u_h = u + \delta u_h, \quad (9.11.1)$$

$$v_{ih} = v_i + \delta v_{ih} \quad (9.11.2)$$

for all $i \in I_A \cup \{0\}$.

- (2) Regarding the shape derivatives of $f_0, f_{i_1}, \dots, f_{i|I_A|}$, we write the numerical solutions of \mathbf{g}_i for each $i \in I_A \cup \{0\}$ in Eq. (9.8.5) obtained using the formulae based on the shape derivative of a function as

$$\mathbf{g}_{ih} = \mathbf{g}_i + \delta \mathbf{g}_{ih}. \quad (9.11.3)$$

Moreover, the numerical solutions of $\bar{\mathbf{g}}_i$ for each $i \in I_A \cup \{0\}$ in Eq. (9.8.35) obtained using the formulae based on the partial shape derivative of a function is written as

$$\bar{\mathbf{g}}_{ih} = \bar{\mathbf{g}}_i + \delta \bar{\mathbf{g}}_{ih}. \quad (9.11.4)$$

Here, \mathbf{g}_i and $\bar{\mathbf{g}}_i$ are functions of $u, v_0, v_{i_1}, \dots, v_{i|I_A|}$ and \mathbf{g}_{ih} and $\bar{\mathbf{g}}_{ih}$ are functions of $u_h, v_{0h}, v_{i_1h}, \dots, v_{i|I_A|h}$, respectively.

- (3) We write the exact solutions of the H^1 gradient method (for example, Problem 9.9.3) calculated using $\mathbf{g}_0, \mathbf{g}_{i_1}, \dots, \mathbf{g}_{i|I_A|}$ as $\varphi_{g0}, \varphi_{gi_1}, \dots, \varphi_{gi|I_A|}$. Moreover, the exact solutions of the H^1 gradient method calculated using $\mathbf{g}_{0h}, \mathbf{g}_{i_1h}, \dots, \mathbf{g}_{i|I_A|h}$ are written as

$$\hat{\varphi}_{gi} = \varphi_{gi} + \delta \hat{\varphi}_{gi} \quad (9.11.5)$$

for all $i \in I_A \cup \{0\}$. With respect to (2), the exact solutions and numerical solutions obtained via the formulae using the partial shape derivative of a function will have (\cdot) attached. The exact solutions of the H^1 gradient method in this case will be written as

$$\hat{\bar{\varphi}}_{gi} = \bar{\varphi}_{gi} + \delta\hat{\bar{\varphi}}_{gi}. \quad (9.11.6)$$

- (4) The numerical solutions of the H^1 gradient method calculated using $\mathbf{g}_{0h}, \mathbf{g}_{i_1h}, \dots, \mathbf{g}_{i_{|I_A|}h}$ are written as

$$\varphi_{gih} = \hat{\bar{\varphi}}_{gi} + \delta\hat{\bar{\varphi}}_{gih} = \varphi_{gi} + \delta\varphi_{gih} \quad (9.11.7)$$

for all $i \in I_A \cup \{0\}$. Moreover, the numerical solutions of the H^1 gradient method obtained using the formulae based on the partial shape derivative of a function are written as

$$\bar{\varphi}_{gih} = \hat{\bar{\varphi}}_{gi} + \delta\hat{\bar{\varphi}}_{gih} = \bar{\varphi}_{gi} + \delta\bar{\varphi}_{gih}. \quad (9.11.8)$$

- (5) The coefficient matrix $(\langle \mathbf{g}_i, \varphi_{gj} \rangle)_{(i,j) \in I_A^2}$ of Eq. (9.10.3) constructed from $\mathbf{g}_0, \mathbf{g}_{i_1}, \dots, \mathbf{g}_{i_{|I_A|}}$ and $\varphi_{g0}, \varphi_{gi_1}, \dots, \varphi_{gi_{|I_A|}}$ is written as \mathbf{A} . Moreover, the coefficient matrix $(\langle \mathbf{g}_{ih}, \varphi_{gjh} \rangle)_{(i,j) \in I_A^2}$ of Eq. (9.10.3) constructed using $\mathbf{g}_{0h}, \mathbf{g}_{i_1h}, \dots, \mathbf{g}_{i_{|I_A|}h}$ and $\varphi_{g0h}, \varphi_{gi_1h}, \dots, \varphi_{gi_{|I_A|}h}$ is written as $\mathbf{A}_h = \mathbf{A} + \delta\mathbf{A}_h$. Furthermore, assuming $f_i = 0$ for all $i \in I_A$, we write $-(\langle \mathbf{g}_i, \varphi_{gj} \rangle)_{i \in I_A}$ as \mathbf{b} . Moreover, the expression $-(\langle \mathbf{g}_{ih}, \varphi_{gjh} \rangle)_{i \in I_A}$ is written as $\mathbf{b}_h = \mathbf{b} + \delta\mathbf{b}_h$. In addition, the exact solutions for the Lagrange multipliers are written as $\boldsymbol{\lambda} = \mathbf{A}^{-1}\mathbf{b}$. On the other hand, its numerical solution is written as

$$\boldsymbol{\lambda}_h = (\lambda_{ih})_{i \in I_A} = \mathbf{A}_h^{-1}\mathbf{b}_h = \boldsymbol{\lambda} + \delta\boldsymbol{\lambda}_h. \quad (9.11.9)$$

Additionally, the exact solutions and numerical solutions obtained using the formulae based on the partial shape derivative of a function will have (\cdot) attached. The numerical solutions for the Lagrange multipliers in this case are written as

$$\bar{\boldsymbol{\lambda}}_h = (\bar{\lambda}_{ih})_{i \in I_A} = \bar{\mathbf{A}}_h^{-1}\bar{\mathbf{b}}_h = \bar{\boldsymbol{\lambda}} + \delta\bar{\boldsymbol{\lambda}}_h. \quad (9.11.10)$$

- (6) Equation (9.10.2) constructed from $\varphi_{g0h}, \varphi_{gi_1h}, \dots, \varphi_{gi_{|I_A|}h}$ and $\lambda_{i_1h}, \dots, \lambda_{i_{|I_A|}h}$ is written as

$$\varphi_{gh} = \varphi_{g0h} + \sum_{i \in I_A} \lambda_{ih} \varphi_{gih} = \varphi_g + \delta\varphi_{gh}. \quad (9.11.11)$$

Moreover, Eq. (9.10.2) obtained using the formulae based on the partial shape derivative of a function is written as

$$\bar{\varphi}_{gh} = \bar{\varphi}_{g0h} + \sum_{i \in I_A} \bar{\lambda}_{ih} \bar{\varphi}_{gih} = \bar{\varphi}_g + \delta \bar{\varphi}_{gh}. \quad (9.11.12)$$

In the above definitions, the error for the search vector is given by $\delta \varphi_{gh}$ and $\delta \bar{\varphi}_{gh}$ of Eqs. (9.11.11) and (9.11.12), respectively. Hence, the aim of this section is to conduct an order evaluation of h with respect to their norms. If such a result can be obtained, the way to select the order of the basis function such that the numerical solution for the search vector converging to the exact solution will be apparent. Here, the following assumptions are essential.

Hypothesis 9.11.1 (Error Estimation of φ_g and $\bar{\varphi}_g$) For $q_R > d$ and $k_1, k_2, j \in \{1, 2, \dots\}$, we assume the following conditions hold:

- (1) The homogeneous forms of the exact solutions u of the state determination problem and $v_0, v_{i_1}, \dots, v_{i_{|I_A|}}$ of the adjoint problem with respect to $f_0, f_{i_1}, \dots, f_{i_{|I_A|}}$ are elements of

$$\mathcal{S} = U \cap W^{\max\{k_1, k_2\}+1, 2q_R} (D; \mathbb{R}). \quad (9.11.13)$$

If necessary, Hypotheses 9.5.1, 9.6.1, 9.5.2, 9.6.2 and 9.5.3 will be amended so that this assumption holds. Also, we let $\partial\Omega$ be of class $H^2 \cap C^{0,1}$ where $\tilde{\Gamma}_0 = \Gamma_{p0} \cup \Gamma_{\eta 00} \cup \Gamma_{\eta 10} \cup \dots \cup \Gamma_{\eta m0}$ belongs to a class of piecewise $H^3 \cap C^{1,1}$, and $X_1 = X \cap W^{1, q_R} (D; \mathbb{R}^d)$ be the linear space of ϕ .

- (2) If the formulae based on the shape derivative of a function are used, the integrands of the cost function f_i for each $i \in I_A \cup \{0\}$ satisfy

$$\zeta_{iu} \nabla u \in L^\infty (D; \mathbb{R}^d), \quad (9.11.14)$$

$$\zeta_i \nabla u (\nabla u)^\top \in L^\infty (D; \mathbb{R}^{d \times d}). \quad (9.11.15)$$

- (3) There exist some positive constants c_1, c_2, c_3 and \bar{c}_3 which do not depend on h such that

$$\|\delta u_h\|_{W^{j, 2q_R}(\Omega; \mathbb{R})} \leq c_1 h^{k_1+1-j} |u|_{W^{k_1+1, 2q_R}(\Omega; \mathbb{R})}, \quad (9.11.16)$$

$$\|\delta v_{ih}\|_{W^{j, 2q_R}(\Omega; \mathbb{R})} \leq c_2 h^{k_1+1-j} |v_i|_{W^{k_1+1, 2q_R}(\Omega; \mathbb{R})}, \quad (9.11.17)$$

$$\|\delta \hat{\varphi}_{gih}\|_{W^{j, 2q_R}(\Omega; \mathbb{R}^d)} \leq c_3 h^{k_2+1-j} |\hat{\varphi}_{gi}|_{W^{k_2+1, 2q_R}(\Omega; \mathbb{R}^d)}, \quad (9.11.18)$$

$$\|\delta \hat{\varphi}_{gih}\|_{W^{j, 2q_R}(\Omega; \mathbb{R}^d)} \leq \bar{c}_3 h^{k_2+1-j} |\hat{\varphi}_{gi}|_{W^{k_2+1, 2q_R}(\Omega; \mathbb{R}^d)} \quad (9.11.19)$$

for all $i \in I_A \cup \{0\}$.

- (4) With respect to the coefficient matrices A_h and \bar{A}_h of Eqs. (9.11.9) and (9.11.10), respectively, there exist positive constants c_4 and \bar{c}_4 that satisfy

$$\|A_h^{-1}\|_{\mathbb{R}^{|I_A| \times |I_A|}} \leq c_4, \quad (9.11.20)$$

$$\|\bar{A}_h^{-1}\|_{\mathbb{R}^{|I_A| \times |I_A|}} \leq \bar{c}_4, \quad (9.11.21)$$

where $\|\cdot\|_{\mathbb{R}^{|I_A| \times |I_A|}}$ represents the norm of the matrix (see Eq. (4.4.3)).

Since $k_1 \in \{1, 2, \dots\}$, Hypothesis 9.11.1 (1) is a stronger condition than \mathcal{S} defined in Eq. (9.5.2). The reason for this is because in Hypothesis 9.11.1 (3), the right-hand side of Eqs. (9.11.17) and (9.11.16) require u and $v_0, v_{i_1}, \dots, v_{i_{|I_A|}}$ to be of class $W^{k_1+1, 2q_R}$. Hypothesis 9.11.1 (3) is based on Corollary 6.6.4. Hypothesis 9.11.1 (4) is a condition which holds when $\mathbf{g}_{i_1}, \dots, \mathbf{g}_{i_{|I_A|}}$ are linearly independent.

We shall give the result of the error estimation in Theorem 9.11.5 after we have proved the following lemmas.

Lemma 9.11.2 (Error Estimation of \mathbf{g}_i and $\bar{\mathbf{g}}_i$) Suppose Hypothesis 9.11.1 (1) and (2) as well as Eqs. (9.11.16) and (9.11.17) are satisfied. Then, there exist positive constants c_5 and \bar{c}_5 which do not depend on h with respect to $\delta\mathbf{g}_{ih}$ and $\delta\bar{\mathbf{g}}_{ih}$ of Eqs. (9.11.3) and (9.11.4), respectively, and the estimates

$$\langle \delta\mathbf{g}_{ih}, \boldsymbol{\varphi} \rangle \leq c_5 h^{k_1-1} \|\boldsymbol{\varphi}\|_{X_1}, \quad (9.11.22)$$

$$\langle \delta\bar{\mathbf{g}}_{ih}, \boldsymbol{\varphi} \rangle \leq \bar{c}_5 h^{k_1-1} \|\boldsymbol{\varphi}\|_{X_1} \quad (9.11.23)$$

hold for all $\boldsymbol{\varphi} \in X_1$. Furthermore, when Hypothesis 9.8.3 (3) is satisfied, we also have the estimate

$$\langle \delta\mathbf{g}_{ih}, \boldsymbol{\varphi} \rangle \leq c_5 h^{k_1} \|\boldsymbol{\varphi}\|_{X_1}. \quad (9.11.24)$$

□

Proof The numerical error $\delta\mathbf{g}_{ih}$ of \mathbf{g}_i using the formulae based on the shape derivative of a function is a numerical error due to δu_h and δv_{ih} . Hence, from Eq. (9.8.5),

$$|\langle \delta\mathbf{g}_{ih}, \boldsymbol{\varphi} \rangle| \leq |\mathcal{L}_{i\phi'uv_i}(\boldsymbol{\phi}, u, v_i) [\boldsymbol{\varphi}, \delta u_h, \delta v_{ih}]| \quad (9.11.25)$$

is established. If the Hölder inequality (Theorem A.9.1), the Poincaré inequality (Corollary A.9.4) and the trace theorem (Theorem 4.4.2) are used, the right-hand

side of Eq. (9.11.25) is suppressed as

$$\begin{aligned}
& |\mathcal{L}_{i\phi'uv_i}(\phi, u, v_i) [\varphi, \delta u_h, \delta v_{ih}]| \\
& \leq \|\delta \mathbf{G}_{\Omega ih}\|_{L^{q_R}(\Omega; \mathbb{R}^{d \times d})} \left\| \nabla \varphi^\top \right\|_{L^2(\Omega; \mathbb{R}^{d \times d})} \\
& \quad + \|\delta g_{\Omega ih}\|_{L^{q_R}(\Omega; \mathbb{R})} \|\nabla \cdot \varphi\|_{L^2(\Omega; \mathbb{R})} \\
& \quad + \|\delta \mathbf{g}_{\zeta bih}\|_{L^{q_R}(\Omega; \mathbb{R}^d)} \|\varphi\|_{L^2(\Omega; \mathbb{R}^d)} \\
& \quad + \|\delta \mathbf{g}_{pih}\|_{L^\infty(\Gamma_p; \mathbb{R}^d)} \|\varphi\|_{L^2(\Gamma_p; \mathbb{R}^d)} \\
& \quad + \|\delta \mathbf{g}_{\partial pih}\|_{L^\infty(\partial \Gamma_p \cup \Theta_p; \mathbb{R}^d)} \|\varphi\|_{L^2(\partial \Gamma_p \cup \Theta_p; \mathbb{R}^d)} \\
& \quad + \|\delta \mathbf{g}_{\eta ih}\|_{L^\infty(\Gamma_{\eta i}; \mathbb{R}^d)} \|\varphi\|_{L^2(\Gamma_{\eta i}; \mathbb{R}^d)} \\
& \quad + \|\delta \mathbf{g}_{\partial \eta ih}\|_{L^\infty(\partial \Gamma_{\eta i} \cup \Theta_{\eta i}; \mathbb{R}^d)} \|\varphi\|_{L^2(\partial \Gamma_{\eta i} \cup \Theta_{\eta i}; \mathbb{R}^d)} \\
& \leq \left\{ \|\delta \mathbf{G}_{\Omega ih}\|_{L^\infty(\Omega; \mathbb{R}^{d \times d})} + \|\delta g_{\Omega ih}\|_{L^\infty(\Omega; \mathbb{R})} + \|\delta \mathbf{g}_{\zeta bih}\|_{L^\infty(\Omega; \mathbb{R}^d)} \right. \\
& \quad \left. + \|\gamma_{\partial \Omega}\| \left(\|\delta \mathbf{g}_{pih}\|_{L^\infty(\Gamma_p; \mathbb{R}^d)} + \|\gamma_{\partial \Gamma_p}\| \|\delta \mathbf{g}_{\partial pih}\|_{L^\infty(\partial \Gamma_p \cup \Theta_p; \mathbb{R}^d)} \right. \right. \\
& \quad \left. \left. + \|\delta \mathbf{g}_{\eta ih}\|_{L^\infty(\Gamma_{\eta i}; \mathbb{R}^d)} + \|\gamma_{\partial \Gamma_{\eta i}}\| \|\delta \mathbf{g}_{\partial \eta ih}\|_{L^\infty(\partial \Gamma_{\eta i} \cup \Theta_{\eta i}; \mathbb{R}^d)} \right) \right\} \|\varphi\|_{X_1}. \tag{9.11.26}
\end{aligned}$$

Here, $\|\gamma_{\partial \Omega}\|$, $\|\gamma_{\partial \Gamma_p}\|$ and $\|\gamma_{\partial \Gamma_{\eta i}}\|$, respectively, represent the norms of the following trace operators for $\varphi \in X_1$:

$$\begin{aligned}
\gamma_{\partial \Omega} & : W^{1, q_R}(\Omega; \mathbb{R}^d) \rightarrow W^{1-1/q_R, q_R}(\partial \Omega; \mathbb{R}^d), \\
\gamma_{\partial \Gamma_p} & : W^{1-1/q_R, q_R}(\partial \Omega; \mathbb{R}^d) \rightarrow W^{1-2/q_R, q_R}(\partial \Gamma_p \cup \Theta_p; \mathbb{R}^d), \\
\gamma_{\partial \Gamma_{\eta i}} & : W^{1-1/q_R, q_R}(\partial \Omega; \mathbb{R}^d) \rightarrow W^{1-2/q_R, q_R}(\partial \Gamma_{\eta i} \cup \Theta_{\eta i}; \mathbb{R}^d)
\end{aligned}$$

and are bounded from the trace theorem because $\partial \Omega$ was assumed to be of class $H^2 \cap C^{0,1}$ in Hypothesis 9.11.1 (1). Moreover, we have the following estimates:

$$\begin{aligned}
& \|\delta \mathbf{G}_{\Omega ih}\|_{L^{q_R}(\Omega; \mathbb{R}^{d \times d})} \\
& \leq 2 \left(\|\nabla \delta u_h\|_{L^{2q_R}(\Omega; \mathbb{R}^d)} \|\nabla v_i\|_{L^{2q_R}(\Omega; \mathbb{R}^d)} \right. \\
& \quad \left. + \|\nabla u\|_{L^{2q_R}(\Omega; \mathbb{R}^d)} \|\nabla \delta v_{ih}\|_{L^{2q_R}(\Omega; \mathbb{R}^d)} \right) \\
& \quad + \|\zeta_i u \nabla u\|_{L^\infty(\Omega; \mathbb{R}^d)} \|\delta u_h\|_{L^{2q_R}(\Omega; \mathbb{R})} \|\nabla u\|_{L^{2q_R}(\Omega; \mathbb{R}^d)}
\end{aligned}$$

$$\begin{aligned}
& + \left\| \zeta_i \nabla u (\nabla u)^\top \right\|_{L^\infty(\Omega; \mathbb{R}^{d \times d})} \|\nabla \delta u_h\|_{L^{2q_R}(\Omega; \mathbb{R}^d)} \|\nabla u\|_{L^{2q_R}(\Omega; \mathbb{R}^d)} \\
& + \left\| \zeta_i (\nabla u)^\top \right\|_{W^{1,2q_R}(\Omega; \mathbb{R}^d)} \|\nabla \delta u_h\|_{L^{2q_R}(\Omega; \mathbb{R}^d)} \\
\leq & 2 \left(\|\delta u_h\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \|v_i\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \right. \\
& \left. + \|u\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \|\delta v_{ih}\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \right) \\
& + \|\zeta_{iu} \nabla u\|_{L^\infty(\Omega; \mathbb{R}^d)} \|\delta u_h\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \|u\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \\
& + \left\| \zeta_i \nabla u (\nabla u)^\top \right\|_{L^\infty(\Omega; \mathbb{R}^{d \times d})} \|\delta u_h\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \|u\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \\
& + \left\| \zeta_i (\nabla u)^\top \right\|_{W^{1,2q_R}(\Omega; \mathbb{R}^d)} \|\delta u_h\|_{W^{1,2q_R}(\Omega; \mathbb{R})}, \tag{9.11.27}
\end{aligned}$$

$$\begin{aligned}
& \|\delta g_{\Omega ih}\|_{L^{q_R}(\Omega; \mathbb{R})} \\
\leq & \|\zeta_{iu}\|_{L^{2q_R}(\Omega; \mathbb{R})} \|\delta u_h\|_{L^{2q_R}(\Omega; \mathbb{R})} \\
& + \left\| \zeta_i (\nabla u)^\top \right\|_{L^{2q_R}(\Omega; \mathbb{R}^d)} \|\nabla \delta u_h\|_{L^{2q_R}(\Omega; \mathbb{R}^d)} \\
& + \|\nabla \delta u_h\|_{L^{2q_R}(\Omega; \mathbb{R}^d)} \|\nabla v_i\|_{L^{2q_R}(\Omega; \mathbb{R}^d)} \\
& + \|\nabla u\|_{L^{2q_R}(\Omega; \mathbb{R}^d)} \|\nabla \delta v_{ih}\|_{L^{2q_R}(\Omega; \mathbb{R}^d)} \\
& + \|b\|_{L^{2q_R}(\Omega; \mathbb{R})} \|\delta v_{ih}\|_{L^{2q_R}(\Omega; \mathbb{R})} \\
\leq & \|\zeta_{iu}\|_{L^{2q_R}(\Omega; \mathbb{R})} \|\delta u_h\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \\
& + \left\| \zeta_i (\nabla u)^\top \right\|_{L^{2q_R}(\Omega; \mathbb{R}^d)} \|\delta u_h\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \\
& + \|\delta u_h\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \|v_i\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \\
& + \|u\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \|\delta v_{ih}\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \\
& + \|b\|_{L^{2q_R}(\Omega; \mathbb{R})} \|\delta v_{ih}\|_{W^{1,2q_R}(\Omega; \mathbb{R})}, \tag{9.11.28}
\end{aligned}$$

$$\begin{aligned}
& \|\delta \mathbf{g}_{\zeta bih}\|_{L^{q_R}(\Omega; \mathbb{R}^d)} \\
\leq & \|b'\|_{L^{2q_R}(\Omega; \mathbb{R})} \|\delta v_{ih}\|_{L^{2q_R}(\Omega; \mathbb{R})} \leq \|b'\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \|\delta v_{ih}\|_{W^{1,2q_R}(\Omega; \mathbb{R})}, \tag{9.11.29}
\end{aligned}$$

$$\begin{aligned}
& \|\delta \mathbf{g}_{pih}\|_{L^\infty(\Gamma_p; \mathbb{R}^d)} \\
\leq & \|\kappa\|_{C^0(\Gamma_p; \mathbb{R})} \|\mathbf{v}\|_{L^\infty(\Gamma_p; \mathbb{R}^d)} \|p_N\|_{L^{2q_R}(\Gamma_p; \mathbb{R})} \|\delta v_{ih}\|_{L^{2q_R}(\Gamma_p; \mathbb{R})} \\
& + (d-1) \max_{i \in \{1, \dots, d-1\}} \|\boldsymbol{\tau}_i\|_{L^\infty(\Gamma_p; \mathbb{R}^d)} \left(\|\nabla p_N\|_{L^{2q_R}(\Gamma_p; \mathbb{R}^d)} \|\delta v_{ih}\|_{L^{2q_R}(\Gamma_p; \mathbb{R})} \right)
\end{aligned}$$

$$\begin{aligned}
& + \|p_N\|_{L^{2q_R}(\Gamma_p; \mathbb{R})} \|\nabla \delta v_{ih}\|_{L^{2q_R}(\Gamma_p; \mathbb{R}^d)} \Big) + \|p'_N\|_{L^{2q_R}(\Gamma_p; \mathbb{R})} \|\delta v_{ih}\|_{L^{2q_R}(\Gamma_p; \mathbb{R})} \\
& \leq \|\kappa\|_{C^0(\Gamma_p; \mathbb{R})} \|\nu\|_{L^\infty(\Gamma_p; \mathbb{R}^d)} \|\gamma_{\partial\Omega}\|^2 \|p_N\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \|\delta v_{ih}\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \\
& + (d-1) \max_{i \in \{1, \dots, d-1\}} \|\tau_i\|_{L^\infty(\Gamma_p; \mathbb{R}^d)} \|\gamma_{\partial\Omega}\|^2 \\
& \times \left(\|p_N\|_{W^{2,2q_R}(\Omega; \mathbb{R})} \|\delta v_{ih}\|_{W^{1,2q_R}(\Omega; \mathbb{R})} + \|p_N\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \|\delta v_{ih}\|_{W^{2,2q_R}(\Omega; \mathbb{R})} \right) \\
& + \|\gamma_{\partial\Omega}\|^2 \|p'_N\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \|\delta v_{ih}\|_{W^{1,2q_R}(\Omega; \mathbb{R})}, \tag{9.11.30}
\end{aligned}$$

$$\begin{aligned}
& \|\delta \mathbf{g}_{\partial pih}\|_{L^\infty(\partial\Gamma_p \cup \Theta_p; \mathbb{R}^d)} \\
& \leq \|\tau\|_{L^\infty(\partial\Gamma_p \cup \Theta_p; \mathbb{R}^d)} \|p_N\|_{L^{2q_R}(\partial\Gamma_p \cup \Theta_p; \mathbb{R})} \|\delta v_{ih}\|_{L^{2q_R}(\partial\Gamma_p \cup \Theta_p; \mathbb{R})} \\
& \leq \|\tau\|_{L^\infty(\partial\Gamma_p \cup \Theta_p; \mathbb{R}^d)} \|\gamma_{\partial\Omega}\|^2 \|\gamma_{\partial\Gamma}\|^2 \|p_N\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \|\delta v_{ih}\|_{W^{1,2q_R}(\Omega; \mathbb{R})}. \tag{9.11.31}
\end{aligned}$$

A similar result is obtained for $\|\delta \mathbf{g}_{\eta ih}\|_{L^\infty(\Gamma_{\eta i}; \mathbb{R}^d)} \|\delta \mathbf{g}_{\partial \eta ih}\|_{L^\infty(\partial\Gamma_{\eta i} \cup \Theta_{\eta i}; \mathbb{R}^d)}$. Here, if Hypothesis 9.11.1 (1) and (2) are satisfied, all the expressions without the terms with δ are bounded. Moreover, if we focus on the terms with δ , there is a term containing $\|\delta v_{ih}\|_{W^{2,2q_R}(\Omega; \mathbb{R}^d)}$ in Eq. (9.11.30). Similarly, $\|\delta u_h\|_{W^{2,2q_R}(\Omega; \mathbb{R})}$ is contained in the inequality equation for $\|\delta \mathbf{g}_{\eta ih}\|_{L^\infty(\Gamma_{\eta i}; \mathbb{R}^d)}$. Hence, if Eqs. (9.11.16) and (9.11.17) with $j = 2$ are substituted in for the terms with δ , these terms become bounded. Hence, we can obtain Eq. (9.11.22).

Furthermore, if Hypothesis 9.8.3 (3) (Eq. (9.8.9) to Eq. (9.8.13) are zero, or $\tilde{\Gamma}_0 = \Gamma_{p0} \cup \Gamma_{\eta i0} \subset \tilde{\Omega}_{C0}$ in Eq. (9.1.1)) is satisfied, since the terms with τ disappear, there are no terms which contain $\|\delta u_h\|_{W^{2,2q_R}(\Omega; \mathbb{R})}$ and $\|\delta v_{ih}\|_{W^{2,2q_R}(\Omega; \mathbb{R}^d)}$. Hence, Eqs. (9.11.16) and (9.11.17) with $j = 1$ can be substituted into the terms with δ to obtain Eq. (9.11.24).

On the other hand, the numerical error $\delta \bar{\mathbf{g}}_{ih}$ using the formulae based on the partial shape derivative of a function satisfies

$$|\langle \delta \bar{\mathbf{g}}_{ih}, \boldsymbol{\varphi} \rangle| \leq |\mathcal{L}_{i\phi^*uv_i}(\boldsymbol{\phi}, u, v_i) [\boldsymbol{\varphi}, \delta u_h, \delta v_{ih}]| \tag{9.11.32}$$

from Eq. (9.8.35). If the Hölder inequality (Theorem A.9.1), the Poincaré inequality (Corollary A.9.4) and the trace theorem (Theorem 4.4.2) are used, the right-hand side of Eq. (9.11.32) is suppressed as

$$\begin{aligned}
& |\mathcal{L}_{i\phi^*uv_i}(\boldsymbol{\phi}, u, v_i) [\boldsymbol{\varphi}, \delta u_h, \delta v_{ih}]| \\
& \leq \|\delta \bar{\mathbf{g}}_{\zeta bih}\|_{L^{q_R}(\Omega; \mathbb{R}^d)} \|\boldsymbol{\varphi}\|_{L^2(\Omega; \mathbb{R}^d)} \\
& + \|\delta \bar{\mathbf{g}}_{\partial\Omega ih}\|_{L^\infty(\partial\Omega; \mathbb{R}^d)} \|\boldsymbol{\varphi}\|_{L^2(\partial\Omega; \mathbb{R}^d)} \\
& + \|\delta \bar{\mathbf{g}}_{pih}\|_{L^\infty(\Gamma_p; \mathbb{R}^d)} \|\boldsymbol{\varphi}\|_{L^2(\Gamma_p; \mathbb{R}^d)}
\end{aligned}$$

$$\begin{aligned}
& + \|\delta \bar{\mathbf{g}}_{\partial pih}\|_{L^\infty(\partial\Gamma_p \cup \Theta_p; \mathbb{R}^d)} \|\boldsymbol{\varphi}\|_{L^2(\partial\Gamma_p \cup \Theta_p; \mathbb{R}^d)} \\
& + \|\delta \bar{\mathbf{g}}_{\eta ih}\|_{L^\infty(\Gamma_{\eta i}; \mathbb{R}^d)} \|\boldsymbol{\varphi}\|_{L^2(\Gamma_{\eta i}; \mathbb{R}^d)} \\
& + \|\delta \bar{\mathbf{g}}_{\partial \eta ih}\|_{L^\infty(\partial\Gamma_{\eta i} \cup \Theta_{\eta i}; \mathbb{R}^d)} \|\boldsymbol{\varphi}\|_{L^2(\partial\Gamma_{\eta i} \cup \Theta_{\eta i}; \mathbb{R}^d)} \\
& + \|\delta \bar{\mathbf{g}}_{Dh}\|_{L^\infty(\Gamma_D; \mathbb{R}^d)} \|\boldsymbol{\varphi}\|_{L^2(\Gamma_D; \mathbb{R}^d)} \\
& \leq \left\{ \|\delta \bar{\mathbf{g}}_{\xi bih}\|_{L^{q_R}(\Omega; \mathbb{R}^d)} + \|\gamma_{\partial\Omega}\|^2 \left(\|\delta \bar{\mathbf{g}}_{\partial\Omega ih}\|_{L^\infty(\partial\Omega; \mathbb{R}^d)} + \|\delta \bar{\mathbf{g}}_{pih}\|_{L^\infty(\Gamma_p; \mathbb{R}^d)} \right. \right. \\
& \quad + \|\gamma_{\partial\Gamma}\| \|\delta \bar{\mathbf{g}}_{\partial pih}\|_{L^\infty(\partial\Gamma_p \cup \Theta_p; \mathbb{R}^d)} + \|\delta \bar{\mathbf{g}}_{\eta ih}\|_{L^\infty(\Gamma_{\eta i}; \mathbb{R}^d)} \\
& \quad \left. \left. + \|\gamma_{\partial\Gamma}\| \|\delta \bar{\mathbf{g}}_{\partial \eta ih}\|_{L^\infty(\partial\Gamma_{\eta i} \cup \Theta_{\eta i}; \mathbb{R}^d)} + \|\delta \bar{\mathbf{g}}_{Dh}\|_{L^\infty(\Gamma_D; \mathbb{R}^d)} \right) \right\} \|\boldsymbol{\varphi}\|_{X_1},
\end{aligned}$$

where

$$\begin{aligned}
\|\delta \bar{\mathbf{g}}_{\xi bih}\|_{L^{q_R}(\Omega; \mathbb{R}^d)} & \leq \|b\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \|\delta v_{ih}\|_{W^{2,2q_R}(\Omega; \mathbb{R})} \\
\|\delta \bar{\mathbf{g}}_{\partial\Omega ih}\|_{L^\infty(\partial\Omega; \mathbb{R}^d)} & \\
\leq & \left(\|\zeta_{iu}\|_{L^{2q_R}(\partial\Omega; \mathbb{R})} \|\delta u_h\|_{L^{2q_R}(\partial\Omega; \mathbb{R})} \right. \\
& + \|\nabla \delta u_h\|_{L^{2q_R}(\partial\Omega; \mathbb{R}^d)} \|\nabla v_i\|_{L^{2q_R}(\partial\Omega; \mathbb{R}^d)} \\
& + \|\nabla u\|_{L^{2q_R}(\partial\Omega; \mathbb{R}^d)} \|\nabla \delta v_{ih}\|_{L^{2q_R}(\partial\Omega; \mathbb{R}^d)} \\
& \left. + \|b\|_{L^{2q_R}(\partial\Omega; \mathbb{R})} \|\delta u_h\|_{L^{2q_R}(\partial\Omega; \mathbb{R})} \right) \|\mathbf{v}\|_{L^\infty(\partial\Omega; \mathbb{R}^d)} \\
\leq & \left(\|\zeta_{iu}\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \|\delta u_h\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \right. \\
& + \|\delta u_h\|_{W^{2,2q_R}(\Omega; \mathbb{R})} \|v_i\|_{W^{2,2q_R}(\Omega; \mathbb{R})} \\
& + \|u\|_{W^{2,2q_R}(\Omega; \mathbb{R})} \|\delta v_{ih}\|_{W^{2,2q_R}(\Omega; \mathbb{R})} \\
& \left. + \|b\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \|\delta u_h\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \right) \|\mathbf{v}\|_{L^\infty(\partial\Omega; \mathbb{R}^d)}, \\
\|\delta \bar{\mathbf{g}}_{pih}\|_{L^\infty(\Gamma_p; \mathbb{R}^d)} & \\
\leq & \|\kappa\|_{C^0(\Gamma_p; \mathbb{R})} \|\mathbf{v}\|_{L^\infty(\Gamma_p; \mathbb{R}^d)} \|p_N\|_{L^{2q_R}(\Gamma_p; \mathbb{R})} \|\delta v_{ih}\|_{L^{2q_R}(\Gamma_p; \mathbb{R})} \\
& + \|\mathbf{v}\|_{L^\infty(\Gamma_p; \mathbb{R}^d)}^2 \left(\|\nabla p_N\|_{L^{2q_R}(\Gamma_p; \mathbb{R}^d)} \|\delta v_{ih}\|_{L^{2q_R}(\Gamma_p; \mathbb{R}^d)} \right. \\
& \quad \left. + \|p_N\|_{L^{2q_R}(\Gamma_p; \mathbb{R})} \|\nabla \delta v_{ih}\|_{L^{2q_R}(\Gamma_p; \mathbb{R}^d)} \right) \\
& + \|p_N^*\|_{L^{2q_R}(\Gamma_p; \mathbb{R})} \|\delta v_{ih}\|_{L^{2q_R}(\Gamma_p; \mathbb{R})}
\end{aligned}$$

$$\begin{aligned}
&\leq \|\kappa\|_{C^0(\Gamma_p; \mathbb{R})} \|\mathbf{v}\|_{L^\infty(\Gamma_p; \mathbb{R}^d)} \|\gamma_{\partial\Omega}\|^2 \|p_N\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \|\delta v_{ih}\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \\
&\quad + \|\mathbf{v}\|_{L^\infty(\Gamma_p; \mathbb{R}^d)}^2 \|\gamma_{\partial\Omega}\|^2 \left(\|p_N\|_{W^{2,2q_R}(\Omega; \mathbb{R})} \|\delta v_{ih}\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \right. \\
&\quad \left. + \|p_N\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \|\delta v_{ih}\|_{W^{2,2q_R}(\Omega; \mathbb{R})} \right) \\
&\quad + \|\gamma_{\partial\Omega}\|^2 \|p_N^*\|_{W^{1,2q_R}(\Omega; \mathbb{R})} \|\delta v_{ih}\|_{W^{1,2q_R}(\Omega; \mathbb{R})}, \\
\|\delta \bar{\mathbf{g}}_{\partial pih}\|_{L^\infty(\partial\Gamma_p \cup \Theta_p; \mathbb{R}^d)} &= \|\delta \mathbf{g}_{\partial pih}\|_{L^\infty(\partial\Gamma_p \cup \Theta_p; \mathbb{R}^d)}, \\
\|\delta \bar{\mathbf{g}}_{Dh}\|_{L^\infty(\Gamma_D; \mathbb{R}^d)} &\leq \|\mathbf{v}\|_{L^\infty(\Gamma_D; \mathbb{R}^d)}^2 \left(\|\nabla \delta u_h\|_{L^{2q_R}(\Gamma_D; \mathbb{R}^d)} \|\nabla v_i\|_{L^{2q_R}(\Gamma_D; \mathbb{R}^d)} \right. \\
&\quad + \|\nabla(u - u_D)\|_{L^{2q_R}(\Gamma_D; \mathbb{R}^d)} \|\nabla \delta v_{ih}\|_{L^{2q_R}(\Gamma_D; \mathbb{R}^d)} \\
&\quad + \|\nabla \delta v_{ih}\|_{L^{2q_R}(\Gamma_D; \mathbb{R}^d)} \|\nabla u\|_{L^{2q_R}(\Gamma_D; \mathbb{R}^d)} \\
&\quad \left. + \|\nabla(v_i - v_{Di})\|_{L^{2q_R}(\Gamma_D; \mathbb{R}^d)} \|\nabla \delta u_h\|_{L^{2q_R}(\Gamma_D; \mathbb{R}^d)} \right) \\
&\leq \|\gamma_{\partial\Omega}\|^2 \|\mathbf{v}\|_{L^\infty(\Gamma_D; \mathbb{R}^d)}^2 \left(\|\delta u_h\|_{W^{2,2q_R}(\Omega; \mathbb{R})} \|v_i\|_{W^{2,2q_R}(\Omega; \mathbb{R})} \right. \\
&\quad + \|(u - u_D)\|_{W^{2,2q_R}(\Omega; \mathbb{R})} \|\delta v_{ih}\|_{W^{2,2q_R}(\Omega; \mathbb{R})} \\
&\quad + \|\delta v_{ih}\|_{W^{2,2q_R}(\Omega; \mathbb{R})} \|u\|_{W^{2,2q_R}(\Omega; \mathbb{R})} \\
&\quad \left. + \|(v_i - v_{Di})\|_{W^{2,2q_R}(\Omega; \mathbb{R})} \|\delta u_h\|_{W^{2,2q_R}(\Omega; \mathbb{R})} \right).
\end{aligned}$$

Similar results can be obtained for $\|\delta \bar{\mathbf{g}}_{\eta ih}\|_{L^\infty(\Gamma_\eta; \mathbb{R}^d)}$ and $\|\delta \bar{\mathbf{g}}_{\partial\eta ih}\|_{L^\infty(\partial\Gamma_\eta \cup \Theta_\eta; \mathbb{R}^d)}$. Here, if Hypothesis 9.11.1 (1) is satisfied, all expressions without the term with δ are bounded. Moreover, if Eqs. (9.11.16) and (9.11.17) with $j = 2$ are substituted into the terms with δ , Eq. (9.11.23) can be obtained, completing the proof of the lemma. \square

Lemma 9.11.3 (Error Estimation of $\boldsymbol{\varphi}_{gi}$ and $\bar{\boldsymbol{\varphi}}_{gi}$) *Suppose Hypothesis 9.11.1 (1), (2) and Eqs. (9.11.16) and (9.11.17) hold. Then, there exist positive constants c_6 and \bar{c}_6 which do not depend on h such that*

$$\|\delta \boldsymbol{\varphi}_{gih}\|_{X_1} \leq c_6 h^{\min\{k_1-1, k_2\}}, \quad (9.11.33)$$

$$\|\delta \bar{\boldsymbol{\varphi}}_{gih}\|_{X_1} \leq \bar{c}_6 h^{\min\{k_1-1, k_2\}} \quad (9.11.34)$$

holds with respect to $\delta \boldsymbol{\varphi}_{gih}$ and $\delta \bar{\boldsymbol{\varphi}}_{gih}$ of Eqs. (9.11.7) and (9.11.8), respectively. Furthermore, if Hypothesis 9.8.3 (3) is satisfied, then we also have

$$\|\delta \boldsymbol{\varphi}_{gih}\|_{X_1} \leq c_6 h^{\min\{k_1, k_2\}}. \quad (9.11.35)$$

\square

Proof When the formulae of the shape derivative of a function are used,

$$\|\delta\boldsymbol{\varphi}_{gih}\|_{X_1} \leq \|\delta\hat{\boldsymbol{\varphi}}_{gi}\|_{X_1} + \|\delta\hat{\boldsymbol{\varphi}}_{gih}\|_{X_1} \quad (9.11.36)$$

holds because of Eqs. (9.11.5) and (9.11.7). Here, $\|\delta\hat{\boldsymbol{\varphi}}_{gi}\|_{X_1}$ shows the error in the exact solution of the H^1 gradient method (see, for example, Problem 9.9.3) caused by $\delta\mathbf{g}_{ih}$ of Lemma 9.11.2. $\|\delta\hat{\boldsymbol{\varphi}}_{gih}\|_{X_1}$ shows the error in the numerical solution of the H^1 gradient method. $\|\delta\hat{\boldsymbol{\varphi}}_{gi}\|_{X_1}$ of Eq. (9.11.36) satisfies

$$a_X(\delta\hat{\boldsymbol{\varphi}}_{gi}, \boldsymbol{\varphi}) = -\langle \delta\mathbf{g}_{ih}, \boldsymbol{\varphi} \rangle$$

for all $\boldsymbol{\varphi} \in X_1$. Hence, if we let $\boldsymbol{\varphi} = \delta\hat{\boldsymbol{\varphi}}_{gi}$,

$$\alpha_X \|\delta\hat{\boldsymbol{\varphi}}_{gi}\|_{X_1}^2 \leq |\langle \delta\mathbf{g}_{ih}, \delta\hat{\boldsymbol{\varphi}}_{gi} \rangle| \quad (9.11.37)$$

holds, where α_X is a positive constant used in Eq. (9.9.1). With respect to $\delta\mathbf{g}_{ih}$ of Eq. (9.11.37), if Eq. (9.11.22) of Lemma 9.11.2 is used,

$$\|\delta\hat{\boldsymbol{\varphi}}_{gi}\|_{X_1} \leq \frac{c_5}{\alpha_X} h^{k_1-1} \quad (9.11.38)$$

is obtained. On the other hand, $\|\delta\hat{\boldsymbol{\varphi}}_{gih}\|_{H^1(\Omega; \mathbb{R}^d)}$ satisfies

$$\|\delta\hat{\boldsymbol{\varphi}}_{gih}\|_{X_1} \leq \|\delta\hat{\boldsymbol{\varphi}}_{gih}\|_{W^{1,2q_R}(\Omega; \mathbb{R}^d)} \leq c_3 h^{k_2} \|\hat{\boldsymbol{\varphi}}_{gi}\|_{W^{k_2+1,2q_R}(\Omega; \mathbb{R}^d)} \quad (9.11.39)$$

in view of Eq. (9.11.18) with $j = 1$. In Eq. (9.11.39), $\|\hat{\boldsymbol{\varphi}}_{gi}\|_{W^{k_2+1,2q_R}(\Omega; \mathbb{R}^d)}$ is bounded. This is because if Hypothesis 9.11.1 (1) is used in the proof of Theorem 9.9.6, then we have that $\hat{\boldsymbol{\varphi}}_{gi} \in W^{k_2+1,\infty}(\Omega; \mathbb{R}^d)$. Hence, if Eqs. (9.11.38) and (9.11.39) are substituted into Eq. (9.11.36), then we obtain Eq. (9.11.33).

Furthermore, if Hypothesis 9.8.3 (3) is satisfied, Eq. (9.11.24) of Lemma 9.11.2 can then be applied to $\delta\mathbf{g}_{ih}$ of Eq. (9.11.37) to get

$$\|\delta\hat{\boldsymbol{\varphi}}_{gi}\|_{X_1} \leq \frac{c_5}{\alpha_X} h^{k_1}. \quad (9.11.40)$$

Here, if Eqs. (9.11.40) and (9.11.39) are substituted into Eq. (9.11.36), then we arrive at Eq. (9.11.35) of the lemma.

If $\delta\mathbf{g}_{ih}$ of Eq. (9.11.37) is changed to $\delta\bar{\mathbf{g}}_{ih}$ and Eq. (9.11.23) of Lemma 9.11.2 is used with respect to $\delta\bar{\mathbf{g}}_{ih}$, then we get Eq. (9.11.34), which finishes the proof of the lemma. \square

Lemma 9.11.4 (Error Estimation of λ_h and $\bar{\lambda}_h$) Suppose Hypothesis 9.11.1 holds. Then, there exist positive constants c_7 \bar{c}_7 which do not depend on h such

that

$$\|\delta\lambda_h\|_{\mathbb{R}^{|I_A|}} \leq c_7 h^{\min\{k_1-1, k_2\}}, \quad (9.11.41)$$

$$\|\delta\bar{\lambda}_h\|_{\mathbb{R}^{|I_A|}} \leq \bar{c}_7 h^{\min\{k_1-1, k_2\}} \quad (9.11.42)$$

hold with respect to λ_h of Eq. (9.11.9) and $\bar{\lambda}_h$ of Eq. (9.11.10), respectively. Furthermore, if Hypothesis 9.8.3 (3) is satisfied, then we also have

$$\|\delta\lambda_h\|_{\mathbb{R}^{|I_A|}} \leq c_7 h^{\min\{k_1, k_2\}}. \quad (9.11.43)$$

Proof When the formulae based on the shape derivative of a function are used, one has

$$\begin{aligned} \delta\lambda_h &= A_h^{-1} (-\delta A_h \lambda + \delta b_h) \\ &= A_h^{-1} \left\{ - \left((\delta g_{ih}, \varphi_{gj})_{(i,j) \in I_A^2} - (g_i, \delta \varphi_{gjh})_{(i,j) \in I_A^2} \right) \lambda \right. \\ &\quad \left. + (\delta g_{ih}, \varphi_{g0})_{i \in I_A} + (g_i, \delta \varphi_{g0h})_{i \in I_A} \right\} \end{aligned}$$

with respect to λ_h of Eq. (9.11.9). Hence, if Eq. (9.11.20) is used, then we get the bound

$$\begin{aligned} \|\delta\lambda_h\|_{\mathbb{R}^{|I_A|}} &\leq c_4 \left(1 + |I_A| \max_{i \in I_A} |\lambda_i| \right) \\ &\quad \times \max_{(i,j) \in I_A \times (I_A \cup \{0\})} (|\langle \delta g_{ih}, \varphi_{gj} \rangle| + |\langle g_i, \delta \varphi_{gjh} \rangle|). \end{aligned} \quad (9.11.44)$$

For $|\langle \delta g_{ih}, \varphi_{gj} \rangle|$ of Eq. (9.11.44), if Eq. (9.11.22) of Lemma 9.11.2 is used,

$$|\langle \delta g_{ih}, \varphi_{gj} \rangle| \leq c_5 h^{k_1-1} \|\varphi_{gj}\|_X \quad (9.11.45)$$

holds. Moreover, we have

$$|\langle g_i, \delta \varphi_{gjh} \rangle| \leq c_6 h^{k_1-1} \|g_i\|_{X'_1} \quad (9.11.46)$$

from Eq. (9.11.33) of Lemma 9.11.3. In Eq. (9.11.46), $\|g_i\|_{X'_1}$ is bounded. This is because from Theorem 9.8.2, $g_i \in X'$ holds, and using $X' \subset X'_1$, $\|g_i\|_{X'_1} \leq \|g_i\|_{X'} < \infty$ is obtained. Here, if Eqs. (9.11.45) and (9.11.46) are substituted into Eq. (9.11.44), then we obtain Eq. (9.11.41) of the lemma.

Furthermore, if Hypothesis 9.8.3 (3) is satisfied, by applying Eq. (9.11.24) of Lemma 9.11.2 to δg_{ih} of Eq. (9.11.45), then we get Eq. (9.11.43) of the lemma.

If δg_{ih} and $\delta\varphi_{gjh}$ of Eqs. (9.11.45) and (9.11.46) are replaced by $\delta\bar{g}_{ih}$ and $\delta\bar{\varphi}_{gjh}$, respectively, then we can apply Theorem 9.8.2 in place of Theorem 9.8.6, and, in addition, applying Eq. (9.11.33) of Lemma 9.11.3 in place of Eq. (9.11.34), we eventually obtain Eq. (9.11.42), completing the proof of the lemma. \square

The following results can be obtained based on these lemmas.

Theorem 9.11.5 (Error Estimation of φ_g and $\bar{\varphi}_g$) *Suppose Hypothesis 9.11.1 holds. Then, there exist positive constants c and \bar{c} which do not depend on h such that*

$$\|\delta\varphi_{gh}\|_{X_1} \leq ch^{\min\{k_1-1, k_2\}}, \quad (9.11.47)$$

$$\|\delta\bar{\varphi}_{gh}\|_{X_1} \leq \bar{c}h^{\min\{k_1-1, k_2\}} \quad (9.11.48)$$

hold with respect to $\delta\varphi_{gh}$ and $\delta\bar{\varphi}_{gh}$ of Eqs. (9.11.11) and (9.11.12), respectively. Furthermore, if Hypothesis 9.8.3 (3) holds, then we also have

$$\|\delta\varphi_{gh}\|_{X_1} \leq ch^{\min\{k_1, k_2\}}. \quad (9.11.49)$$

Proof From Eq. (9.11.11), we have

$$\delta\varphi_{gh} = \delta\varphi_{g0h} + \sum_{i \in I_A} (\delta\lambda_{ih}\varphi_{gi} + \lambda_i \delta\varphi_{gih}) \quad (9.11.50)$$

from which we get

$$\begin{aligned} \|\delta\varphi_{gh}\|_{X_1} &\leq \left(1 + |I_A| \max_{i \in I_A} |\lambda_i|\right) \max_{i \in I_A \cup \{0\}} \|\delta\varphi_{gih}\|_{X_1} \\ &\quad + \|\delta\lambda_h\|_{\mathbb{R}^{|I_A|}} \max_{i \in I_A} \|\varphi_{gi}\|_{X_1}. \end{aligned} \quad (9.11.51)$$

If Eq. (9.11.33) of Lemma 9.11.3 and Eq. (9.11.41) of Lemma 9.11.4 are substituted into Eq. (9.11.51), Eq. (9.11.47) of the theorem can be obtained.

Furthermore, if Hypothesis 9.8.3 (3) holds, then, by substituting Eq. (9.11.35) of Lemma 9.11.3 and Eq. (9.11.43) of Lemma 9.11.4 into Eq. (9.11.51), we obtain Eq. (9.11.49) of the theorem.

If $\delta\varphi_{gih}$ and $\delta\lambda_h$ of Eq. (9.11.51) are replaced by $\delta\bar{\varphi}_{gih}$ and $\delta\bar{\lambda}_h$, respectively, and Eq. (9.11.34) of Lemma 9.11.3 and Eq. (9.11.42) of Lemma 9.11.4 are substituted into Eq. (9.11.51), then we obtain Eq. (9.11.48) of the theorem, which finishes the proof. \square

Theorem 9.11.5 allows us to infer the following remark about the error estimation of the finite element solution with respect to the shape optimization problem of domain variation type.

Remark 9.11.6 (Error Estimation of φ_g and $\bar{\varphi}_g$) From Theorem 9.11.5, in order to reduce the error $\|\delta\varphi_{gh}\|_{X_1}$ of the search vector φ_{gh} with respect to h of the finite element division sequence $\{\mathcal{T}_h\}_{h \rightarrow 0}$ linearly, the following conditions need to be satisfied.

When Hypothesis 9.11.1 is satisfied:

- (1) use the finite element solutions of the state determination problem and the adjoint problems for $f_0, f_{i_1}, \dots, f_{i_{|I_A|}}$ based on a $k_1 = 2$ -order basis function, and
- (2) use finite element solutions with respect to $\mathbf{g}_0, \mathbf{g}_{i_1}, \dots, \mathbf{g}_{i_{|I_A|}}$ of Eq. (9.8.5) (formulae based on the shape derivative of a function) or $\bar{\mathbf{g}}_0, \bar{\mathbf{g}}_{i_1}, \dots, \bar{\mathbf{g}}_{i_{|I_A|}}$ of Eq. (9.8.35) (formulae based on the partial shape derivative of a function) in the H^1 gradient method based on a $k_2 = 1$ -order basis function.

Furthermore, if Hypothesis 9.8.3 (3) is satisfied:

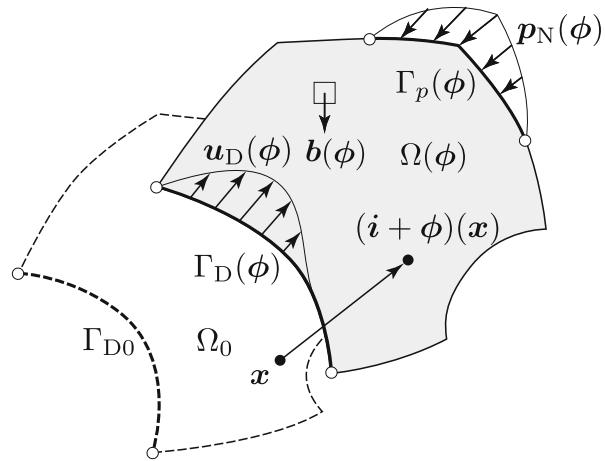
- (1) use the finite element solutions of the state determination problem and the adjoint problems for $f_0, f_{i_1}, \dots, f_{i_{|I_A|}}$ based on a $k_1 = 1$ -order basis function, and
- (2) use the finite element solutions with respect to $\mathbf{g}_0, \mathbf{g}_{i_1}, \dots, \mathbf{g}_{i_{|I_A|}}$ of Eq. (9.8.5) (formulae based on the shape derivative of a function) in the H^1 gradient method based on a $k_2 = 1$ -order basis function. \square

9.12 Shape Optimization Problem of Linear Elastic Body

As an application of the shape optimization problem, let us consider a mean compliance minimization problem of a linear elastic body, and compute the shape derivatives of cost functions associated with the problem. Here too, the conditions with respect to the initial domain Ω_0 , the definitions of Γ_{D0} , Γ_{N0} and Γ_{p0} as well as the definitions of X and \mathcal{D} are taken to be the same as in Sect. 9.1 (Fig. 9.15). However, we describe \mathcal{D} more specifically as follows:

$$\mathcal{D} = \left\{ \boldsymbol{\phi} \in Y \mid \begin{cases} |\boldsymbol{\phi}|_{C^{0,1}(D; \mathbb{R}^d)} \leq \sigma, \\ \|\boldsymbol{\phi}\|_{H^2 \cap C^{0,1}(D; \mathbb{R}^d)} \leq \beta \quad (\Gamma_{p0} = \emptyset \text{ or } \Gamma_{p0} \subset \bar{\Omega}_{C0}), \\ \|\boldsymbol{\phi}\|_{H^3 \cap C^{1,1}(D; \mathbb{R}^d)} \leq \beta \quad (\Gamma_{p0} \not\subset \bar{\Omega}_{C0}) \end{cases} \right\}. \quad (9.12.1)$$

Fig. 9.15 Initial domain $\Omega_0 \subset D$ and domain variation (displacement) ϕ in a linear elastic body



9.12.1 State Determination Problem

Define a linear elastic problem as a state determination problem. In the sequel, the notation of Problem 5.4.2 will be used, and in addition, the precise shape optimization problem will be presented. For a given $\phi \in \mathcal{D}$, let the linear space U of state variable (solution of state determination problem) u be

$$U = \left\{ u \in H^1(D; \mathbb{R}^d) \mid u = \mathbf{0}_{\mathbb{R}^d} \text{ on } \Gamma_D(\phi) \right\}. \quad (9.12.2)$$

Notice that the range of U in Eq. (9.5.1) is \mathbb{R} , but it is \mathbb{R}^d in Eq. (9.12.2). Moreover, the admissible set containing u is taken to be

$$\mathcal{S} = U \cap W^{2,4}(D; \mathbb{R}^d). \quad (9.12.3)$$

In this section too, the required regularity conditions will be specified when necessary. In order to satisfy the regularity requirements, we assume the same set of hypotheses (Hypotheses 9.5.1 and 9.5.2) with respect to the regularities of known functions, where the domain is changed to D and the functions are now denoted by bold letters. In addition, for the linear elastic problems, we let $E(u)$ and $S(\phi, u) = C(\phi)E(u)$ be the linear strain and stress, respectively, that were defined in Eq. (5.4.2) and Eq. (5.4.6). Also, we assume that the stiffness C is elliptic (Eq. (5.4.8)) and bounded (Eq. (5.4.9)). Suppose that in the modified hypotheses, Hypotheses 9.5.1 and 9.5.2 shown above, the condition

$$C \in C_{S'}^1(B; C^{0,1}(D; \mathbb{R}^{d \times d \times d \times d})) \quad (9.12.4)$$

is added. For the regularity of the boundary, Hypothesis 9.5.3 is used.

Using the above assumptions, a linear elastic problem of domain variation type is defined as follows.

Problem 9.12.1 (Linear Elastic Problem of Domain Variation Type) For a $\phi \in \mathcal{D}$, let $\mathbf{b}(\phi)$, $\mathbf{p}_N(\phi)$, $\mathbf{u}_D(\phi)$ and $\mathbf{C}(\phi)$ be given. Find the $\mathbf{u} : \Omega(\phi) \rightarrow \mathbb{R}^d$ which satisfies

$$-\nabla^\top \mathbf{S}(\phi, \mathbf{u}) = \mathbf{b}^\top(\phi) \quad \text{in } \Omega(\phi), \quad (9.12.5)$$

$$\mathbf{S}(\phi, \mathbf{u}) \mathbf{v} = \mathbf{p}_N(\phi) \quad \text{on } \Gamma_p(\phi), \quad (9.12.6)$$

$$\mathbf{S}(\phi, \mathbf{u}) \mathbf{v} = \mathbf{0}_{\mathbb{R}^d} \quad \text{on } \Gamma_N(\phi) \setminus \bar{\Gamma}_p(\phi), \quad (9.12.7)$$

$$\mathbf{u} = \mathbf{u}_D(\phi) \quad \text{on } \Gamma_D(\phi). \quad (9.12.8)$$

□

From now on, we write $\mathbf{S}(\phi, \mathbf{u})$ as $\mathbf{S}(\mathbf{u})$ for simplicity. For later use, referring to the weak form (Problem 5.4.3) of a linear elastic problem and the Dirichlet boundary condition, we define the Lagrange function with respect to Problem 9.12.1 as

$$\begin{aligned} \mathcal{L}_S(\phi, \mathbf{u}, \mathbf{v}) &= \int_{\Omega(\phi)} (-\mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}) + \mathbf{b} \cdot \mathbf{v}) \, dx + \int_{\Gamma_p(\phi)} \mathbf{p}_N \cdot \mathbf{v} \, d\gamma \\ &\quad + \int_{\Gamma_D(\phi)} \{(\mathbf{u} - \mathbf{u}_D) \cdot (\mathbf{S}(\mathbf{v}) \mathbf{v}) + \mathbf{v} \cdot (\mathbf{S}(\mathbf{u}) \mathbf{v})\} \, d\gamma, \end{aligned} \quad (9.12.9)$$

where \mathbf{u} is not necessarily the solution of Problem 9.12.1, and \mathbf{v} is an element of U introduced as a Lagrange multiplier. In this case, if \mathbf{u} is the solution of Problem 9.12.1, then

$$\mathcal{L}_S(\phi, \mathbf{u}, \mathbf{v}) = 0$$

holds for any $\mathbf{v} \in U$. This equation is equivalent to the weak form of Problem 9.12.1.

9.12.2 Mean Compliance Minimization Problem

Let us define a shape optimization problem of linear elastic body. The cost functions we consider here is defined as follows. With respect to the solution \mathbf{u} of Problem 9.12.1, the functional

$$\begin{aligned} f_0(\phi, \mathbf{u}) &= \hat{l}(\phi)(\mathbf{u}) \\ &= \int_{\Omega(\phi)} \mathbf{b} \cdot \mathbf{u} \, dx + \int_{\Gamma_p(\phi)} \mathbf{p}_N \cdot \mathbf{u} \, d\gamma \\ &\quad - \int_{\Gamma_D(\phi)} \mathbf{u}_D \cdot (\mathbf{S}(\mathbf{u}) \mathbf{v}) \, d\gamma \end{aligned} \quad (9.12.10)$$

is referred to as the mean compliance. The reason for such use of the terminology is given in Sect. 8.9.2. Here, $\hat{l}(\phi)(\mathbf{u})$ shows that $\hat{l}(\mathbf{u})$ defined in Eq. (5.2.3) also depends on ϕ . Moreover,

$$f_1(\phi) = \int_{\Omega(\phi)} dx - c_1 \quad (9.12.11)$$

is called a constraint function with respect to the domain measure. Here, c_1 is a positive constant such that $f_1(\phi) \leq 0$ holds with respect to some $\phi \in \mathcal{D}$.

Here, a mean compliance minimization problem is defined as follows.

Problem 9.12.2 (Mean Compliance Minimization Problem) Suppose \mathcal{D} and \mathcal{S} is defined as Eqs. (9.12.1) and (9.12.3), respectively. Let f_0 and f_1 be Eqs. (9.12.10) and (9.12.11). In this case, find $\Omega(\phi)$ such that

$$\min_{(\phi, \mathbf{u} - \mathbf{u}_D) \in \mathcal{D} \times \mathcal{S}} \{f_0(\phi, \mathbf{u}) \mid f_1(\phi) \leq 0, \text{ Problem 9.12.1}\}. \quad \square$$

9.12.3 Shape Derivatives of Cost Functions

Let us obtain the shape derivatives of $f_0(\phi, \mathbf{u})$ and $f_1(\phi)$. Here, we will look separately at the case when the formulae based on the shape derivative of a function is used and the case when the formulae based on the partial shape derivative of a function is utilized. When the formulae based on the shape derivative of a function is used, the corresponding expression up to the second-order shape derivative will be established. As preparation for this, let the Lagrange function of $f_0(\phi, \mathbf{u})$ be

$$\begin{aligned} \mathcal{L}_0(\phi, \mathbf{u}, \mathbf{v}_0) &= f_0(\phi, \mathbf{u}) + \mathcal{L}_S(\phi, \mathbf{u}, \mathbf{v}_0) \\ &= \int_{\Omega(\phi)} (-\mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0) + \mathbf{b} \cdot (\mathbf{u} + \mathbf{v}_0)) dx + \int_{\Gamma_p(\phi)} \mathbf{p}_N \cdot (\mathbf{u} + \mathbf{v}_0) d\gamma \\ &\quad + \int_{\Gamma_D(\phi)} \{(\mathbf{u} - \mathbf{u}_D) \cdot (\mathbf{S}(\mathbf{v}_0) \mathbf{v}) + (\mathbf{v}_0 - \mathbf{u}_D) \cdot (\mathbf{S}(\mathbf{u}) \mathbf{v})\} d\gamma. \end{aligned} \quad (9.12.12)$$

Here, \mathcal{L}_S is the Lagrange function of the state determination problem defined by Eq. (9.12.9). Moreover, \mathbf{v}_0 is the Lagrange multiplier with respect to the state determination problem prepared for f_0 , and $\tilde{\mathbf{v}}_0 = \mathbf{v}_0 - \mathbf{u}_D$ is assumed to be an element of U .

Shape Derivatives of f_0 and f_1 Using Formulae Based on Shape Derivative of a Function

Let us obtain the shape derivative of f_0 using the formulae based on the shape derivative of a function. Here, $\mathbf{b}(\phi)$, $\mathbf{p}_N(\phi)$, $\mathbf{u}_D(\phi)$ and $\mathbf{C}(\phi)$ are assumed to be fixed with the material. Here, if $\mathbf{b}(\phi)$ is written as \mathbf{b} , ϕ is also omitted in other equations.

Here, the Fréchet derivative of \mathcal{L}_0 can be written as

$$\begin{aligned}\mathcal{L}'_0(\phi, \mathbf{u}, \mathbf{v}_0) [\boldsymbol{\varphi}, \hat{\mathbf{u}}, \hat{\mathbf{v}}_0] &= \mathcal{L}_{0\phi'}(\phi, \mathbf{u}, \mathbf{v}_0) [\boldsymbol{\varphi}] + \mathcal{L}_{0\mathbf{u}}(\phi, \mathbf{u}, \mathbf{v}_0) [\hat{\mathbf{u}}] \\ &\quad + \mathcal{L}_{0\mathbf{v}_0}(\phi, \mathbf{u}, \mathbf{v}_0) [\hat{\mathbf{v}}_0]\end{aligned}\quad (9.12.13)$$

with respect to an arbitrary variation $(\boldsymbol{\varphi}, \hat{\mathbf{u}}, \hat{\mathbf{v}}_0) \in X \times U \times U$. Here, it will go along with the notations of Eqs. (9.3.5) and (9.3.15). Each term is considered below.

The third term on the right-hand side of Eq. (9.12.13) can be rewritten as

$$\mathcal{L}_{0\mathbf{v}_0}(\phi, \mathbf{u}, \mathbf{v}_0) [\hat{\mathbf{v}}_0] = \mathcal{L}_{S\mathbf{v}_0}(\phi, \mathbf{u}, \mathbf{v}_0) [\hat{\mathbf{v}}_0] = \mathcal{L}_S(\phi, \mathbf{u}, \hat{\mathbf{v}}_0). \quad (9.12.14)$$

Equation (9.12.14) is the Lagrange function of the state determination problem (Problem 9.12.1). Hence, if \mathbf{u} is the weak solution of the state determination problem, the third term on the right-hand side of Eq. (9.12.13) equates to zero.

Moreover, the second term on the right-hand side of Eq. (9.12.13) can be written as

$$\begin{aligned}\mathcal{L}_{0\mathbf{u}}(\phi, \mathbf{u}, \mathbf{v}_0) [\hat{\mathbf{u}}] &= \int_{\Omega(\phi)} (-S(\hat{\mathbf{u}}) \cdot \mathbf{E}(\mathbf{v}_0) + \mathbf{b} \cdot \hat{\mathbf{u}}) \, dx + \int_{\Gamma_p(\phi)} \mathbf{p}_N \cdot \hat{\mathbf{u}} \, d\gamma \\ &\quad + \int_{\Gamma_D(\phi)} \{\hat{\mathbf{u}} \cdot (S(\mathbf{v}_0) \mathbf{v}) + (\mathbf{v}_0 - \mathbf{u}) \cdot (S(\hat{\mathbf{u}}) \mathbf{v})\} \, d\gamma \\ &= \mathcal{L}_S(\phi, \mathbf{v}_0, \hat{\mathbf{u}}).\end{aligned}\quad (9.12.15)$$

If Eqs. (9.12.15) and (9.12.14) are compared, it is clear that it is a relationship whereby \mathbf{v}_0 and \mathbf{u} are swapped over. Hence, if the self-adjoint relationship

$$\mathbf{v}_0 = \mathbf{u} \quad (9.12.16)$$

holds, the second term on the right-hand side of Eq. (9.12.13) vanishes.

Furthermore, the first term on the right-hand side of Eq. (9.12.13) becomes

$$\begin{aligned}\mathcal{L}_{0\phi'}(\phi, \mathbf{u}, \mathbf{v}_0) [\boldsymbol{\varphi}] &= \int_{\Omega(\phi)} \left[\left(S(\mathbf{u}) (\nabla \mathbf{v}_0^\top)^\top + S(\mathbf{v}_0) (\nabla \mathbf{u}^\top)^\top \right) \cdot \nabla \boldsymbol{\varphi}^\top \right.\end{aligned}$$

$$\begin{aligned}
& + \{-S(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0) + \mathbf{b} \cdot (\mathbf{u} + \mathbf{v}_0)\} \nabla \cdot \boldsymbol{\varphi} \Big] dx \\
& + \int_{\Gamma_p(\boldsymbol{\phi})} [\kappa \{ \mathbf{p}_N \cdot (\mathbf{u} + \mathbf{v}_0) \} \mathbf{v} \cdot \boldsymbol{\varphi} - \nabla_\tau \{ \mathbf{p}_N \cdot (\mathbf{u} + \mathbf{v}_0) \} \cdot \boldsymbol{\varphi}_\tau] d\gamma \\
& + \int_{\partial\Gamma_p(\boldsymbol{\phi}) \cup \Theta(\boldsymbol{\phi})} \{ \mathbf{p}_N \cdot (\mathbf{u} + \mathbf{v}_0) \} \boldsymbol{\tau} \cdot \boldsymbol{\varphi} d\zeta \\
& + \int_{\Gamma_D(\boldsymbol{\phi})} [\{(\mathbf{u} - \mathbf{u}_D) \cdot \mathbf{w}(\boldsymbol{\varphi}, \mathbf{v}_0) + (\mathbf{v}_0 - \mathbf{u}_D) \cdot \mathbf{w}(\boldsymbol{\varphi}, \mathbf{u})\} \\
& \quad + \{(\mathbf{u} - \mathbf{u}_D) \cdot (S(\mathbf{v}_0) \mathbf{v}) + (\mathbf{v}_0 - \mathbf{u}_D) \cdot (S(\mathbf{u}) \mathbf{v})\} (\nabla \cdot \boldsymbol{\varphi})_\tau] d\gamma
\end{aligned} \tag{9.12.17}$$

using Eq. (9.3.5), representing the result of Proposition 9.3.4, and Eq. (9.3.15) of Proposition 9.3.7, where

$$\mathbf{w}(\boldsymbol{\varphi}, \mathbf{u}) = S(\mathbf{u}) \left[\left\{ \mathbf{v} \cdot (\nabla \boldsymbol{\varphi}^\top \mathbf{v}) \right\} \mathbf{v} - \left\{ (\nabla \boldsymbol{\varphi}^\top + (\nabla \boldsymbol{\varphi}^\top)^\top) \mathbf{v} \right\} \mathbf{v} \right] \tag{9.12.18}$$

and $(\nabla \cdot \boldsymbol{\varphi})_\tau$ follows Eq. (9.2.6). In order to obtain Eq. (9.12.17), the following identity:

$$\begin{aligned}
& - (S(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0))_{\boldsymbol{\phi}'} [\boldsymbol{\varphi}] \\
& = - (\mathbf{E}(\mathbf{u}) \cdot S(\mathbf{v}_0))_{\boldsymbol{\phi}'} [\boldsymbol{\varphi}] \\
& = (\mathbf{E}(\mathbf{u}) \cdot S(\mathbf{v}_0))_{\nabla \mathbf{u}^\top} \cdot (\nabla \boldsymbol{\varphi}^\top \nabla \mathbf{u}^\top) \\
& \quad + (S(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0))_{\nabla \mathbf{v}_0^\top} \cdot (\nabla \boldsymbol{\varphi}^\top \nabla \mathbf{v}_0^\top) - S(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0) (\nabla \cdot \boldsymbol{\varphi}) \\
& = (\nabla \boldsymbol{\varphi}^\top \nabla \mathbf{u}^\top)^s \cdot S(\mathbf{v}_0) + S(\mathbf{u}) \cdot (\nabla \boldsymbol{\varphi}^\top \nabla \mathbf{v}_0^\top)^s \\
& \quad - S(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0) (\nabla \cdot \boldsymbol{\varphi}) \\
& = (\nabla \boldsymbol{\varphi}^\top \nabla \mathbf{u}^\top) \cdot S(\mathbf{v}_0) + S(\mathbf{u}) \cdot (\nabla \boldsymbol{\varphi}^\top \nabla \mathbf{v}_0^\top) \\
& \quad - S(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0) (\nabla \cdot \boldsymbol{\varphi}) \\
& = \left(S(\mathbf{u}) (\nabla \mathbf{v}_0^\top)^\top \right) \cdot \nabla \boldsymbol{\varphi}^\top + \left(S(\mathbf{v}_0) (\nabla \mathbf{u}^\top)^\top \right) \cdot \nabla \boldsymbol{\varphi}^\top \\
& \quad - S(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0) (\nabla \cdot \boldsymbol{\varphi})
\end{aligned}$$

is used, which is derived using Eq. (9.8.18). The notation $(\cdot)^s$ represents $((\cdot)^\top + (\cdot))/2$.

With the above results in mind, assume that \mathbf{u} is the weak solution of Problem 9.12.1 and that the self-adjoint relationship (Eq. (9.12.16)) holds. In this case, from the fact that Dirichlet condition holds for Problem 9.12.1, Eq. (9.12.17) can be written as

$$\begin{aligned}\tilde{f}'_0(\boldsymbol{\phi})[\boldsymbol{\varphi}] &= \mathcal{L}_{0\boldsymbol{\phi}'}(\boldsymbol{\phi}, \mathbf{u}, \mathbf{v}_0)[\boldsymbol{\varphi}] = \langle \mathbf{g}_0, \boldsymbol{\varphi} \rangle \\ &= \int_{\Omega(\boldsymbol{\phi})} \left(\mathbf{G}_{\Omega 0} \cdot \nabla \boldsymbol{\varphi}^\top + g_{\Omega 0} \nabla \cdot \boldsymbol{\varphi} \right) dx \\ &\quad + \int_{\Gamma_p(\boldsymbol{\phi})} \mathbf{g}_{p0} \cdot \boldsymbol{\varphi} dy + \int_{\partial\Gamma_p(\boldsymbol{\phi}) \cup \Theta(\boldsymbol{\phi})} \mathbf{g}_{\partial p0} \cdot \boldsymbol{\varphi} d\zeta \quad (9.12.19)\end{aligned}$$

using the notation of Eq. (7.5.15) for \tilde{f}_0 , where

$$\mathbf{G}_{\Omega 0} = 2\mathbf{S}(\mathbf{u}) \left(\nabla \mathbf{u}^\top \right)^\top, \quad (9.12.20)$$

$$g_{\Omega 0} = -\mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{u}) + 2\mathbf{b} \cdot \mathbf{u}, \quad (9.12.21)$$

$$\mathbf{g}_{p0} = 2\kappa (\mathbf{p}_N \cdot \mathbf{u}) \mathbf{v}, \quad (9.12.22)$$

$$\mathbf{g}_{\partial p0} = 2 (\mathbf{p}_N \cdot \mathbf{u}) \boldsymbol{\tau}. \quad (9.12.23)$$

From the results above, similar conclusions with Theorem 9.8.2 can be obtained for \mathbf{g}_0 of Eq. (9.12.19).

On the other hand, the shape derivative of $f_1(\boldsymbol{\phi})$ is obtained as

$$f'_1(\boldsymbol{\phi})[\boldsymbol{\varphi}] = \langle \mathbf{g}_1, \boldsymbol{\varphi} \rangle = \int_{\Omega(\boldsymbol{\phi})} g_{\Omega 1} \nabla \cdot \boldsymbol{\varphi} dx, \quad (9.12.24)$$

where

$$g_{\Omega 1} = 1. \quad (9.12.25)$$

This is established by letting $u = 1$ in Proposition 9.3.1 without using \mathcal{L}_S , which, on the other hand, is due to the fact that the solution to the state determination problem is not used.

Second-Order Shape Derivatives of f_0 and f_1 Using Formulae Based on Shape Derivative of a Function

Now, let us obtain the second-order shape derivatives of the mean compliance f_0 and the constraint cost function f_1 with respect to the domain measure of linear elastic body. Here, the formulae based on the shape derivative of a function is used following the procedures shown in Sect. 9.8.2.

Firstly, let us think about the second-order shape derivative of f_0 . We assume that $\mathbf{b} = \mathbf{0}_{\mathbb{R}^d}$ corresponding to Hypothesis 9.8.3 (1). The relationship corresponding to Hypothesis 9.8.3 (2) is satisfied here. Moreover, assume (3) in Hypothesis 9.8.3.

The Lagrange function \mathcal{L}_0 of f_0 is defined by Eq. (9.12.12). Viewing (ϕ, \mathbf{u}) as a design variable, we define its corresponding admissible set and admissible direction set as

$$S = \{(\phi, \mathbf{u}) \in \mathcal{D} \times \mathcal{S} \mid \mathcal{L}_S(\phi, \mathbf{u}, \mathbf{v}) = 0 \text{ for all } \mathbf{v} \in U\},$$

$$T_S(\phi, \mathbf{u}) = \{(\varphi, \hat{\mathbf{v}}) \in X \times U \mid \mathcal{L}_{S\phi\mathbf{u}}(\phi, \mathbf{u}, \mathbf{v})[\varphi, \hat{\mathbf{v}}] = 0 \text{ for all } \mathbf{v} \in U\}.$$

Considering Eq. (9.1.6), the second-order Fréchet partial derivative of \mathcal{L}_0 of Eq. (9.12.12) with respect to arbitrary variations $(\varphi_1, \hat{\mathbf{v}}_1), (\varphi_2, \hat{\mathbf{v}}_2) \in T_S(\phi, \mathbf{u})$ of the design variable $(\phi, \mathbf{u}) \in S$ becomes

$$\begin{aligned} & \mathcal{L}_{0(\phi', \mathbf{u})(\phi', \mathbf{u})}(\phi, \mathbf{u}, \mathbf{v}_0)[(\varphi_1, \hat{\mathbf{v}}_1), (\varphi_2, \hat{\mathbf{v}}_2)] \\ &= \left(\mathcal{L}_{0(\phi', \mathbf{u})} \right)_{(\phi', \mathbf{u})}(\phi, \mathbf{u}, \mathbf{v}_0)[(\varphi_1, \hat{\mathbf{v}}_1), (\varphi_2, \hat{\mathbf{v}}_2)] + \langle \mathbf{g}_0(\phi), \mathbf{t}(\varphi_1, \varphi_2) \rangle \\ &= (\mathcal{L}_{0\phi'}(\phi, \mathbf{u}, \mathbf{v}_0)[\varphi_1] + \mathcal{L}_{0\mathbf{u}}(\phi, \mathbf{u}, \mathbf{v}_0)[\hat{\mathbf{v}}_1])_{\phi'}[\varphi_2] \\ &+ (\mathcal{L}_{0\phi'}(\phi, \mathbf{u}, \mathbf{v}_0)[\varphi_1] + \mathcal{L}_{0\mathbf{u}}(\phi, \mathbf{u}, \mathbf{v}_0)[\hat{\mathbf{v}}_1])_{\mathbf{u}}[\hat{\mathbf{v}}_2] \\ &+ \langle \mathbf{g}_0(\phi), \mathbf{t}(\varphi_1, \varphi_2) \rangle \\ &= (\mathcal{L}_{0\phi'}(\phi, \mathbf{u}, \mathbf{v}_0)[\varphi_1, \varphi_2] + \mathcal{L}_{0\phi'\mathbf{u}}(\phi, \mathbf{u}, \mathbf{v}_0)[\varphi_1, \hat{\mathbf{v}}_2] \\ &+ \mathcal{L}_{0\phi'\mathbf{u}}(\phi, \mathbf{u}, \mathbf{v}_0)[\varphi_2, \hat{\mathbf{v}}_1] + \mathcal{L}_{0\mathbf{u}\mathbf{u}}(\phi, \mathbf{u}, \mathbf{v}_0)[\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2] \\ &+ \langle \mathbf{g}_0(\phi), \mathbf{t}(\varphi_1, \varphi_2) \rangle), \end{aligned} \quad (9.12.26)$$

where $\langle \mathbf{g}_0(\phi), \mathbf{t}(\varphi_1, \varphi_2) \rangle$ follows the definition given in Eq. (9.1.8).

Here, the first and fifth terms in Eq. (9.12.26) becomes

$$\begin{aligned} & (\mathcal{L}_{0\phi'}(\phi, \mathbf{u}, \mathbf{v}_0)[\varphi_1, \varphi_2] + \langle \mathbf{g}_0(\phi), \mathbf{t}(\varphi_1, \varphi_2) \rangle) \\ &= \int_{\Omega(\phi)} \left[\left\{ \left(\mathbf{S}(\mathbf{u}) \left(\nabla \mathbf{v}_0^\top \right)^\top \right) \cdot \nabla \varphi_1^\top \right\}_{\phi'}[\varphi_2] \right. \\ &+ \left\{ \left(\mathbf{S}(\mathbf{v}_0) \left(\nabla \mathbf{u}^\top \right)^\top \right) \cdot \nabla \varphi_1^\top \right\}_{\phi'}[\varphi_2] \\ &- (\mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0)) (\nabla \cdot \varphi_1)_{\phi'}[\varphi_2] \\ &+ 2 \mathbf{S}(\mathbf{u}) \left(\nabla \mathbf{u}^\top \right)^\top \cdot \left(\nabla \varphi_2^\top \nabla \varphi_1^\top - \nabla \varphi_1^\top (\nabla \cdot \varphi_2) \right) \\ &\left. - \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{u}) \left\{ \left(\nabla \varphi_2^\top \right)^\top \cdot \nabla \varphi_1^\top - (\nabla \cdot \varphi_2) (\nabla \cdot \varphi_1) \right\} \right] dx. \end{aligned} \quad (9.12.27)$$

Here, Eq. (9.3.11) was used. The first term of the integrand on the right-hand side of Eq. (9.12.27) becomes

$$\begin{aligned}
& \left[\left\{ S(\mathbf{u}) (\nabla \mathbf{v}_0^\top)^\top \right\} \cdot \nabla \boldsymbol{\varphi}_1^\top \right]_{\boldsymbol{\phi}'} [\boldsymbol{\varphi}_2] \\
&= \left\{ S(\mathbf{u}) \cdot (\nabla \boldsymbol{\varphi}_1^\top \nabla \mathbf{v}_0^\top) \right\}_{\boldsymbol{\phi}'} [\boldsymbol{\varphi}_2] \\
&= \left\{ E(\mathbf{u}) \cdot \left(C (\nabla \boldsymbol{\varphi}_1^\top \nabla \mathbf{v}_0^\top)^s \right) \right\}_{\boldsymbol{\phi}'} [\boldsymbol{\varphi}_2] \\
&= \left[\nabla \mathbf{v}_0^\top \cdot \left\{ (\nabla \boldsymbol{\varphi}_1^\top)^\top S(\mathbf{u}) \right\} \right]_{\boldsymbol{\phi}'} [\boldsymbol{\varphi}_2] \\
&= - \left\{ E(\mathbf{u}) \cdot \left(C (\nabla \boldsymbol{\varphi}_1^\top \nabla \mathbf{v}_0^\top)^s \right) \right\}_{\nabla \mathbf{u}^\top} \cdot (\nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{u}^\top) \\
&\quad - \left[\left\{ S(\mathbf{u}) (\nabla \mathbf{v}_0^\top)^\top \right\} \cdot \nabla \boldsymbol{\varphi}_1^\top \right]_{\nabla \boldsymbol{\varphi}_1^\top} \cdot (\nabla \boldsymbol{\varphi}_2^\top \nabla \boldsymbol{\varphi}_1^\top) \\
&\quad - \left[\nabla \mathbf{v}_0^\top \cdot \left\{ (\nabla \boldsymbol{\varphi}_1^\top)^\top S(\mathbf{u}) \right\} \right]_{\nabla \mathbf{v}_0^\top} \cdot (\nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{v}_0^\top) \\
&\quad + \left[\left\{ S(\mathbf{u}) (\nabla \mathbf{v}_0^\top)^\top \right\} \cdot \nabla \boldsymbol{\varphi}_1^\top \right] \nabla \cdot \boldsymbol{\varphi}_2 \\
&= - (\nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{u}^\top)^s \cdot \left(C (\nabla \boldsymbol{\varphi}_1^\top \nabla \mathbf{v}_0^\top)^s \right) \\
&\quad - \left\{ S(\mathbf{u}) (\nabla \mathbf{v}_0^\top)^\top \right\} \cdot (\nabla \boldsymbol{\varphi}_2^\top \nabla \boldsymbol{\varphi}_1^\top) \\
&\quad - (\nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{v}_0^\top) \cdot \left\{ (\nabla \boldsymbol{\varphi}_1^\top)^\top S(\mathbf{u}) \right\} \\
&\quad + \left[\left\{ S(\mathbf{u}) (\nabla \mathbf{v}_0^\top)^\top \right\} \cdot \nabla \boldsymbol{\varphi}_1^\top \right] \nabla \cdot \boldsymbol{\varphi}_2. \tag{9.12.28}
\end{aligned}$$

Similarly, the second term of the integrand on the right-hand side of Eq. (9.12.27) is similar to Eq. (9.12.28) with \mathbf{u} and \mathbf{v}_0 interchanged. The third term of the integrand on the right-hand side of Eq. (9.12.27) is

$$\begin{aligned}
& - (S(\mathbf{u}) \cdot E(\mathbf{v}_0)) \{ \nabla \cdot \boldsymbol{\varphi}_1 \}_{\boldsymbol{\phi}'} [\boldsymbol{\varphi}_2] \\
&= - S(\mathbf{u}) \cdot E(\mathbf{v}_0) \left\{ - (\nabla \boldsymbol{\varphi}_2^\top)^\top \cdot \nabla \boldsymbol{\varphi}_1^\top + (\nabla \cdot \boldsymbol{\varphi}_2) (\nabla \cdot \boldsymbol{\varphi}_1) \right\}. \tag{9.12.29}
\end{aligned}$$

Hence, noting the self-adjoint relationship, Eq. (9.12.27) becomes

$$\begin{aligned}
& (\mathcal{L}_{0\phi'}(\phi, \mathbf{u}, \mathbf{v}_0)[\varphi_1, \varphi_2] + \langle \mathbf{g}_0(\phi), \mathbf{t}(\varphi_1, \varphi_2) \rangle) \\
&= \int_{\Omega(\phi)} \left[- \left(\nabla \varphi_2^\top \nabla \mathbf{u}^\top \right)^s \cdot \left(\mathbf{C} \left(\nabla \varphi_1^\top \nabla \mathbf{v}_0^\top \right)^s \right) \right. \\
&\quad - \left(\nabla \varphi_2^\top \nabla \mathbf{v}_0^\top \right) \cdot \left\{ \left(\nabla \varphi_1^\top \right)^\top \mathbf{S}(\mathbf{u}) \right\} \\
&\quad - \left(\nabla \varphi_2^\top \nabla \mathbf{v}_0^\top \right)^s \cdot \left(\mathbf{C} \left(\nabla \varphi_1^\top \nabla \mathbf{u}^\top \right)^s \right) \\
&\quad \left. - \left(\nabla \varphi_2^\top \nabla \mathbf{u}^\top \right) \cdot \left\{ \left(\nabla \varphi_1^\top \right)^\top \mathbf{S}(\mathbf{v}_0) \right\} \right] dx \quad (9.12.30)
\end{aligned}$$

Next, consider the second term on the right-hand side of Eq. (9.12.26). If Eq. (9.12.17) with the Dirichlet condition of the state determination problem substituted in is used, we get

$$\begin{aligned}
& \mathcal{L}_{0\phi' \mathbf{u}}(\phi, \mathbf{u}, \mathbf{v}_0)[\varphi_1, \hat{\mathbf{v}}_2] \\
&= \int_{\Omega(\phi)} \left[\left\{ \mathbf{S}(\hat{\mathbf{v}}_2) \left(\nabla \mathbf{v}_0^\top \right)^\top + \mathbf{S}(\mathbf{v}_0) \left(\nabla \hat{\mathbf{v}}_2^\top \right)^\top \right\} \cdot \nabla \varphi_1^\top \right. \\
&\quad \left. - (\mathbf{S}(\hat{\mathbf{v}}_2) \cdot \mathbf{E}(\mathbf{v}_0)) \nabla \cdot \varphi_1 \right] dx. \quad (9.12.31)
\end{aligned}$$

On the other hand, the variation of \mathbf{u} satisfying the state determination problem with respect to an arbitrary domain variation $\varphi_j \in Y$ for $j \in \{1, 2\}$ is written as $\hat{\mathbf{v}}_j = \mathbf{v}'(\phi)[\varphi_j]$. If the Fréchet partial derivative of the Lagrange function \mathcal{L}_S of the state determination problem is taken, we obtain

$$\begin{aligned}
& \mathcal{L}_{S\phi' \mathbf{u}}(\phi, \mathbf{u}, \mathbf{v})[\varphi_j, \hat{\mathbf{v}}_j] \\
&= \int_{\Omega(\phi)} \left\{ \mathbf{S}(\mathbf{u}) \cdot \left(\nabla \varphi_j^\top \nabla \mathbf{v}^\top \right)^s + \mathbf{S}(\mathbf{v}) \cdot \left(\nabla \varphi_j^\top \nabla \mathbf{u}^\top \right)^s \right. \\
&\quad \left. - (\mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v})) \nabla \cdot \varphi_j - \mathbf{S}(\hat{\mathbf{v}}_j) \cdot \mathbf{E}(\mathbf{v}) \right\} dx \\
&= \int_{\Omega(\phi)} \left[\left\{ \left(\nabla \varphi_j^\top \right)^\top \mathbf{S}(\mathbf{u}) + \mathbf{C} \left(\nabla \varphi_j^\top \nabla \mathbf{u}^\top \right)^s - \mathbf{S}(\mathbf{u}) \nabla \cdot \varphi_j \right. \right. \\
&\quad \left. \left. - \mathbf{S}(\hat{\mathbf{v}}_j) \left\{ \left(\nabla \mathbf{v}^\top \right)^\top \right\} \cdot \mathbf{I} \right] dx
\end{aligned}$$

$$\begin{aligned}
&= \int_{\Omega(\phi)} \left[\nabla \mathbf{v}^\top \mathbf{S}(\mathbf{u}) \nabla \boldsymbol{\varphi}_j^\top \right. \\
&\quad \left. + \mathbf{S}(\mathbf{v}) \left\{ \left(\nabla \mathbf{u}^\top \right)^\top \left(\left(\nabla \boldsymbol{\varphi}_j^\top \right)^\top - \nabla \cdot \boldsymbol{\varphi}_j \right) - \left(\nabla \hat{\mathbf{v}}_j^\top \right)^\top \right\} \right] \cdot \mathbf{I} \, dx \\
&= 0
\end{aligned} \tag{9.12.32}$$

for any $\mathbf{v} \in U$. Here, the Dirichlet boundary conditions of \mathbf{v} and $\hat{\mathbf{v}}_j$ were used. From the fact that Eq. (9.12.32) holds with respect to an arbitrary $\mathbf{v} \in U$, the identities

$$\begin{aligned}
&\mathbf{S}(\hat{\mathbf{v}}_j) \cdot \mathbf{E}(\mathbf{v}) \\
&= \mathbf{S}(\mathbf{u}) \cdot \left(\nabla \boldsymbol{\varphi}_j^\top \nabla \mathbf{v}^\top \right)^s + \mathbf{S}(\mathbf{v}) \cdot \left(\nabla \boldsymbol{\varphi}_j^\top \nabla \mathbf{u}^\top \right)^s - (\mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v})) \nabla \cdot \boldsymbol{\varphi}_j,
\end{aligned} \tag{9.12.33}$$

$$\begin{aligned}
&\mathbf{S}(\hat{\mathbf{v}}_j) \left(\nabla \mathbf{v}^\top \right)^\top \\
&= \left\{ \left(\nabla \boldsymbol{\varphi}_j^\top \right)^\top \mathbf{S}(\mathbf{u}) + \mathbf{C} \left(\nabla \boldsymbol{\varphi}_j^\top \nabla \mathbf{u}^\top \right)^s - \nabla \cdot \boldsymbol{\varphi}_j \mathbf{S}(\mathbf{u}) \right\} \left(\nabla \mathbf{v}^\top \right)^\top
\end{aligned} \tag{9.12.34}$$

$$\begin{aligned}
&\mathbf{S}(\mathbf{v}) \left(\nabla \hat{\mathbf{v}}_j^\top \right)^\top \\
&= \nabla \mathbf{v}^\top \mathbf{S}(\mathbf{u}) \nabla \boldsymbol{\varphi}_j^\top + \mathbf{S}(\mathbf{v}) \left(\nabla \mathbf{u}^\top \right)^\top \left\{ \left(\nabla \boldsymbol{\varphi}_j^\top \right)^\top - \nabla \cdot \boldsymbol{\varphi}_j \right\}
\end{aligned} \tag{9.12.35}$$

are obtained. Substituting from Eq. (9.12.33) to Eq. (9.12.35) into Eq. (9.12.31), we get

$$\begin{aligned}
&\mathcal{L}_{0\phi' u}(\phi, \mathbf{u}, \mathbf{v}_0) [\boldsymbol{\varphi}_1, \hat{\mathbf{v}}_2] \\
&= \int_{\Omega(\phi)} \left[\left\{ \left(\nabla \boldsymbol{\varphi}_2^\top \right)^\top \mathbf{S}(\mathbf{u}) + \mathbf{C} \left(\nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{u}^\top \right)^s - \nabla \cdot \boldsymbol{\varphi}_2 \mathbf{S}(\mathbf{u}) \right\} \left(\nabla \mathbf{v}_0^\top \right)^\top \right. \\
&\quad \left. + \nabla \mathbf{v}_0^\top \mathbf{S}(\mathbf{u}) \nabla \boldsymbol{\varphi}_2^\top + \mathbf{S}(\mathbf{v}_0) \left(\nabla \mathbf{u}^\top \right)^\top \left(\left(\nabla \boldsymbol{\varphi}_2^\top \right)^\top - \nabla \cdot \boldsymbol{\varphi}_2 \right) \right\} \cdot \nabla \boldsymbol{\varphi}_1^\top \\
&\quad - \left\{ \mathbf{S}(\mathbf{u}) \cdot \left(\nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{v}_0^\top \right)^s + \mathbf{S}(\mathbf{v}_0) \cdot \left(\nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{u}^\top \right)^s \right. \\
&\quad \left. - (\mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0)) \nabla \cdot \boldsymbol{\varphi}_2 \right\} \nabla \cdot \boldsymbol{\varphi}_1 \right] dx.
\end{aligned} \tag{9.12.36}$$

Similarly, the third term on the right-hand side of Eq. (9.12.26) takes the form as in Eq. (9.12.36) with $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$ interchanged. Lastly, the fourth term on the right-hand side of Eq. (9.12.26) is actually equal to zero.

Summarizing the above results, the second-order shape derivative of \tilde{f}_0 becomes

$$\begin{aligned}
 h_0(\boldsymbol{\phi}, \mathbf{u}, \mathbf{u})[\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2] &= \int_{\Omega(\boldsymbol{\phi})} \left[2S(\mathbf{u}) \cdot \mathbf{E}(\mathbf{u}) (\nabla \cdot \boldsymbol{\varphi}_2) (\nabla \cdot \boldsymbol{\varphi}_1) \right. \\
 &\quad + \left(S(\mathbf{u}) (\nabla \mathbf{u}^\top)^\top \right) \cdot \left\{ \nabla \boldsymbol{\varphi}_2^\top \nabla \boldsymbol{\varphi}_1^\top + \nabla \boldsymbol{\varphi}_1^\top \nabla \boldsymbol{\varphi}_2^\top + \nabla \boldsymbol{\varphi}_2^\top (\nabla \boldsymbol{\varphi}_1^\top)^\top \right. \\
 &\quad \left. \left. + \nabla \boldsymbol{\varphi}_1^\top (\nabla \boldsymbol{\varphi}_2^\top)^\top \right\} - 4 \nabla \boldsymbol{\varphi}_2^\top \nabla \cdot \boldsymbol{\varphi}_1 - 4 \nabla \boldsymbol{\varphi}_1^\top \nabla \cdot \boldsymbol{\varphi}_2 \right] dx. \tag{9.12.37}
 \end{aligned}$$

On the other hand, the second-order shape derivative of $f_1(\boldsymbol{\phi})$ becomes

$$h_1(\boldsymbol{\phi})[\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2] = (f'_1)'(\boldsymbol{\phi}) + \langle \mathbf{g}_1(\boldsymbol{\phi}), \mathbf{t}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) \rangle = 0 \tag{9.12.38}$$

with respect to arbitrary variations $\boldsymbol{\varphi}_1 \in Y$ and $\boldsymbol{\varphi}_2 \in Y$. Here, Eq. (9.3.11) was used.

Second-Order Shape Derivative of Cost Function Using Lagrange Multiplier Method

The application of the Lagrange multiplier method in obtaining the second-order shape derivative of the mean compliance f_0 is described as follows. Fixing $\boldsymbol{\varphi}_1$, we define the Lagrange function for $\tilde{f}'_0(\boldsymbol{\phi})[\boldsymbol{\varphi}_1] = \langle \mathbf{g}_0, \boldsymbol{\varphi}_1 \rangle$ in Eq. (9.12.19) by

$$\mathcal{L}_{10}(\boldsymbol{\phi}, \mathbf{u}, \mathbf{w}_0) = \langle \mathbf{g}_0, \boldsymbol{\varphi}_1 \rangle + \mathcal{L}_S(\boldsymbol{\phi}, \mathbf{u}, \mathbf{w}_0), \tag{9.12.39}$$

where \mathcal{L}_S is given by Eq. (9.12.9), and $\mathbf{w}_0 \in U$ is the adjoint variable provided for \mathbf{u} in \mathbf{g}_0 .

Considering Eq. (9.1.6), with respect to arbitrary variations $(\boldsymbol{\varphi}_2, \hat{\mathbf{u}}, \hat{\mathbf{w}}_0) \in \mathcal{D} \times U^2$ of $(\boldsymbol{\phi}, \mathbf{u}, \mathbf{w}_0)$, the Fréchet derivative of \mathcal{L}_{10} is written as

$$\begin{aligned}
 \mathcal{L}'_{10}(\boldsymbol{\phi}, \mathbf{u}, \mathbf{w}_0)[\boldsymbol{\varphi}_2, \hat{\mathbf{u}}, \hat{\mathbf{w}}_0] &= \mathcal{L}_{10\boldsymbol{\phi}'}(\boldsymbol{\phi}, \mathbf{u}, \mathbf{w}_0)[\boldsymbol{\varphi}_2] + \langle \mathbf{g}_0(\boldsymbol{\phi}), \mathbf{t}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) \rangle + \mathcal{L}_{10\mathbf{u}}(\boldsymbol{\phi}, \mathbf{u}, \mathbf{w}_0)[\hat{\mathbf{u}}] \\
 &\quad + \mathcal{L}_{10\mathbf{w}_0}(\boldsymbol{\phi}, \mathbf{u}, \mathbf{w}_0)[\hat{\mathbf{w}}_0]. \tag{9.12.40}
 \end{aligned}$$

The fourth term on the right-hand side of Eq. (9.12.40) vanishes if \mathbf{u} is the solution of the state determination problem.

Assuming that φ_1 is an H^2 class function in the neighborhood of $\Gamma_p(\phi)$ and then applying Proposition 9.3.7, the third term on the right-hand side of Eq. (9.12.40) is obtained as

$$\begin{aligned} & \mathcal{L}_{10u}(\phi, u, w_0)[\hat{u}] \\ &= \int_{\Omega(\phi)} \left[2 \left\{ \mathbf{C} \left(\nabla \varphi_1^\top \nabla v_0^\top \right)^s + \left(\left(\nabla \varphi_1^\top \right)^\top - \nabla \cdot \varphi_1 \right) S(v_0) \right\} \cdot \nabla \hat{u}^\top \right. \\ & \quad \left. + 2 (\nabla \cdot \varphi_1) \mathbf{b} \cdot \hat{u} - S(w_0) \cdot \mathbf{E}(\hat{u}) \right] dx \\ & \quad + \int_{\Gamma_p(\phi)} (\nabla \cdot \varphi_1)_\tau p_N \cdot \hat{u} d\gamma. \end{aligned} \quad (9.12.41)$$

Here, the condition that Eq. (9.12.41) is zero for arbitrary $\hat{u} \in U$ is equivalent to setting w_0 to be the solution of the following adjoint problem.

Problem 9.12.3 (Adjoint Problem of w_0 with Respect to $\langle g_0, \varphi_1 \rangle$) Under the assumption of Problem 9.12.1, let $\varphi_1 \in Y$ be given. Find $w_0 = w_0(\varphi_1) \in U$ satisfying

$$\begin{aligned} -\nabla^\top S(w_0) &= -2\nabla^\top \left\{ \mathbf{C} \left(\nabla \varphi_1^\top \nabla v_0^\top \right)^s + \left(\left(\nabla \varphi_1^\top \right)^\top - \nabla \cdot \varphi_1 \right) S(v_0) \right\} \\ & \quad + 2\mathbf{b}^\top (\nabla \cdot \varphi_1) \quad \text{in } \Omega(\phi), \\ S(w_0) \mathbf{v} &= (\nabla \cdot \varphi_1)_\tau p_N \quad \text{on } \Gamma_p(\phi), \\ S(w_0) \mathbf{v} &= \mathbf{0}_{\mathbb{R}^d} \quad \text{on } \Gamma_N(\phi) \setminus \bar{\Gamma}_p(\phi), \\ w_0 &= \mathbf{0}_{\mathbb{R}^d} \text{ on } \Gamma_D. \end{aligned}$$

□

Finally, the first and second terms on the right-hand side of Eq. (9.12.40) become

$$\begin{aligned} & \mathcal{L}_{10\phi'}(\phi, u, v_0, w_0(\varphi_1), z_0(\varphi_1))[\varphi_2] + \langle g_0(\phi), t(\varphi_1, \varphi_2) \rangle \\ &= \mathcal{L}_{0\phi'\phi'}(\phi, u, v_0)[\varphi_1, \varphi_2] + \langle g_0(\phi), t(\varphi_1, \varphi_2) \rangle \\ & \quad + \mathcal{L}_{S\phi'}(\phi, u, w_0)[\varphi_2] \end{aligned} \quad (9.12.42)$$

with respect an arbitrary $\varphi_1 \in Y$. The first and second terms on the right-hand side of Eq. (9.12.42) are given by Eq. (9.12.30). The third term is given by

$$\begin{aligned} & \mathcal{L}_{S\phi'}(\phi, u, w_0)[\varphi_2] \\ &= \int_{\Omega(\phi)} \left\{ S(u) \cdot \left(\nabla \varphi_2^\top \nabla w_0^\top \right)^s + S(w_0) \cdot \left(\nabla \varphi_2^\top \nabla u^\top \right)^s \right\} dx \end{aligned}$$

$$\begin{aligned}
& + (\mathbf{b} \cdot \mathbf{w}_0 - \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{w}_0)) \nabla \cdot \boldsymbol{\varphi}_2 \Big] dx \\
& + \int_{\Gamma_p(\boldsymbol{\phi})} \{ \mathbf{p}_N \cdot \mathbf{u} (\nabla \cdot \boldsymbol{\varphi}_1)_\tau + \mathbf{p}_N \cdot \mathbf{w}_0 (\boldsymbol{\varphi}_1) (\boldsymbol{\varphi}_1) \} (\nabla \cdot \boldsymbol{\varphi}_2)_\tau d\gamma.
\end{aligned}$$

Here, \mathbf{u} and $\mathbf{w}_0(\boldsymbol{\varphi}_1)$ are assumed to be the weak solutions of Problems 9.12.1 and 9.12.3, respectively. If we denote $f_0(\boldsymbol{\phi}, \mathbf{u})$ by $\tilde{f}_0(\boldsymbol{\phi})$, then we arrive at the relation

$$\begin{aligned}
& \mathcal{L}_{10\boldsymbol{\phi}'}(\boldsymbol{\phi}, \boldsymbol{\phi}_1, \mathbf{u}, \mathbf{w}_0(\boldsymbol{\varphi}_1)) [\boldsymbol{\varphi}_2] + \langle \mathbf{g}_0(\boldsymbol{\phi}), \mathbf{t}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) \rangle \\
& = \tilde{f}_0''(\boldsymbol{\phi}) [\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2] = \langle \mathbf{g}_{H0}(\boldsymbol{\phi}, \boldsymbol{\varphi}_1), \boldsymbol{\varphi}_2 \rangle \\
& = \int_{\Omega(\boldsymbol{\phi})} \left[-2 \left(\nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{u}^\top \right)^s \cdot \left(\mathbf{C} \left(\nabla \boldsymbol{\varphi}_1^\top \nabla \mathbf{u}^\top \right)^s \right) \right. \\
& \quad \left. - 2 \left\{ \mathbf{S}(\mathbf{u}) \left(\nabla \mathbf{u}^\top \right)^\top \right\} \cdot \left(\nabla \boldsymbol{\varphi}_1^\top \nabla \boldsymbol{\varphi}_2^\top \right) \right. \\
& \quad \left. + \left\{ \mathbf{S}(\mathbf{u}) \left(\nabla \mathbf{w}_0^\top(\boldsymbol{\varphi}_1) \right)^\top + \mathbf{S}(\mathbf{w}_0(\boldsymbol{\varphi}_1)) \left(\nabla \mathbf{u}^\top \right)^\top \right\} \cdot \nabla \boldsymbol{\varphi}_2^\top \right. \\
& \quad \left. - \{ \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{w}_0(\boldsymbol{\varphi}_1)) - \mathbf{b} \cdot \mathbf{w}_0(\boldsymbol{\varphi}_1) \} \nabla \cdot \boldsymbol{\varphi}_2 \right] dx \\
& \quad + \int_{\Gamma_p(\boldsymbol{\phi})} 2 \mathbf{p}_N \cdot (\mathbf{u} + \mathbf{w}_0(\boldsymbol{\varphi}_1)) (\nabla \cdot \boldsymbol{\varphi}_1)_\tau (\nabla \cdot \boldsymbol{\varphi}_2)_\tau d\gamma, \tag{9.12.43}
\end{aligned}$$

where \mathbf{g}_{H0} is the Hesse gradient of the mean compliance.

If $\mathbf{b} = \mathbf{0}_{\mathbb{R}^d}$ and (3) in Hypothesis 9.8.3 are satisfied, with respect to the solution \mathbf{w}_0 of Problem 9.12.3,

$$\begin{aligned}
\mathbf{S}(\mathbf{w}_0) & = 2 \left\{ \mathbf{C} \left(\nabla \boldsymbol{\varphi}_1^\top \nabla \mathbf{v}_0^\top \right)^s \right. \\
& \quad \left. + \left(\left(\nabla \boldsymbol{\varphi}_1^\top \right)^s - \nabla \cdot \boldsymbol{\varphi}_1 \right) \mathbf{S}(\mathbf{v}_0) \right\}, \tag{9.12.44}
\end{aligned}$$

$$\mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{w}_0) = \mathbf{E}(\mathbf{u}) \cdot \mathbf{S}(\mathbf{w}_0), \tag{9.12.45}$$

$$\begin{aligned}
\mathbf{S}(\mathbf{u}) \left(\nabla \mathbf{w}_0^\top \right)^\top \cdot \nabla \boldsymbol{\varphi}_2^\top & = \mathbf{S}(\mathbf{u}) \cdot \left(\nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{w}_0^\top \right)^s \\
& = \left(\nabla \boldsymbol{\varphi}_j^\top \right)^\top \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{w}_0) \\
& = \left(\nabla \boldsymbol{\varphi}_j^\top \right)^\top \mathbf{E}(\mathbf{u}) \cdot \mathbf{S}(\mathbf{w}_0) \tag{9.12.46}
\end{aligned}$$

holds. Here, we used

$$\mathbf{S}(\mathbf{u}) \cdot \left(\nabla \boldsymbol{\varphi}_j^\top \nabla \mathbf{v}_0^\top \right)^s = \left(\nabla \boldsymbol{\varphi}_j^\top \right)^\top \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0).$$

Indeed, this relation is obtained from the fact that the inner product of $\mathbf{S}(\hat{\mathbf{v}}_j)$ obtained from Eq. (9.12.34) and $\mathbf{E}(\mathbf{v})$ accords with Eq. (9.12.33). Substituting Eq. (9.12.44) to Eq. (9.12.46) into Eq. (9.12.43), it can be confirmed that Eq. (9.12.43) accords with Eq. (9.12.37).

Shape Derivatives of f_0 and f_1 Using Formulae Based on Partial Shape Derivative of a Function

Next, let us compute the shape derivative of f_0 using the formulae based on the partial shape derivative of a function. Here, it is assumed that \mathbf{b} , \mathbf{p}_N , \mathbf{u}_D and \mathbf{C} are functions fixed in space. Moreover, we assume that \mathbf{u} and \mathbf{v}_0 are elements of $W^{2,2q_R}(D; \mathbb{R}^d)$ where $q_R > d$.

Under these assumptions, the Fréchet derivative of $\mathcal{L}_0(\boldsymbol{\varphi}, \mathbf{u}, \mathbf{v}_0)$ can be written as

$$\begin{aligned} \mathcal{L}'_0(\boldsymbol{\varphi}, \mathbf{u}, \mathbf{v}_0) [\boldsymbol{\varphi}, \hat{\mathbf{u}}, \hat{\mathbf{v}}_0] &= \mathcal{L}_{0\boldsymbol{\varphi}^*}(\boldsymbol{\varphi}, \mathbf{u}, \mathbf{v}_0) [\boldsymbol{\varphi}] + \mathcal{L}_{0\mathbf{u}}(\boldsymbol{\varphi}, \mathbf{u}, \mathbf{v}_0) [\hat{\mathbf{u}}] \\ &\quad + \mathcal{L}_{0\mathbf{v}_0}(\boldsymbol{\varphi}, \mathbf{u}, \mathbf{v}_0) [\hat{\mathbf{v}}_0] \end{aligned} \quad (9.12.47)$$

for any $(\boldsymbol{\varphi}, \hat{\mathbf{u}}, \hat{\mathbf{v}}_0) \in X \times U \times U$. Here, the notations of Eq. (9.3.21) and Eq. (9.3.27) were used. Each term is considered below.

The third term on the right-hand side of Eq. (9.12.47) is given by Eq. (9.12.14). Hence, if \mathbf{u} is the weak solution of the state determination problem, the said expression equates to zero. Similarly, the second term on the right-hand side of Eq. (9.12.47) is the same as Eq. (9.12.15). Hence, when the self-adjoint relationship (Eq. (9.12.16)) holds, the term also vanishes.

Furthermore, the first term on the right-hand side of Eq. (9.12.47) becomes

$$\begin{aligned} &\mathcal{L}_{0\boldsymbol{\varphi}^*}(\boldsymbol{\varphi}, \mathbf{u}, \mathbf{v}_0) [\boldsymbol{\varphi}] \\ &= \int_{\partial\Omega(\boldsymbol{\varphi})} \{-\mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}_0) + \mathbf{b} \cdot (\mathbf{u} + \mathbf{v}_0)\} \mathbf{v} \cdot \boldsymbol{\varphi} \, d\gamma \\ &\quad + \int_{\Gamma_p(\boldsymbol{\varphi})} (\partial_v + \kappa) \{ \mathbf{p}_N \cdot (\mathbf{u} + \mathbf{v}_0) \} \mathbf{v} \cdot \boldsymbol{\varphi} \, d\gamma \\ &\quad + \int_{\partial\Gamma_p(\boldsymbol{\varphi}) \cup \Theta(\boldsymbol{\varphi})} \{ \mathbf{p}_N \cdot (\mathbf{u} + \mathbf{v}_0) \} \boldsymbol{\tau} \cdot \boldsymbol{\varphi} \, d\zeta \\ &\quad + \int_{\Gamma_D(\boldsymbol{\varphi})} \left[\{(\mathbf{u} - \mathbf{u}_D) \cdot \bar{\mathbf{w}}(\boldsymbol{\varphi}, \mathbf{v}_0) + (\mathbf{v}_0 - \mathbf{u}_D) \cdot \bar{\mathbf{w}}(\boldsymbol{\varphi}, \mathbf{u})\} \right. \end{aligned}$$

$$\begin{aligned}
& + \{(\mathbf{u} - \mathbf{u}_D) \cdot (\mathbf{S}(\mathbf{v}_0) \mathbf{v}) + (\mathbf{v}_0 - \mathbf{u}_D) \cdot (\mathbf{S}(\mathbf{u}) \mathbf{v})\} (\nabla \cdot \boldsymbol{\varphi})_\tau \\
& + (\partial_v + \kappa) \{(\mathbf{u} - \mathbf{u}_D) \cdot (\mathbf{S}(\mathbf{v}_0) \mathbf{v}) + (\mathbf{v}_0 - \mathbf{u}_D) \cdot (\mathbf{S}(\mathbf{u}) \mathbf{v})\} \mathbf{v} \cdot \boldsymbol{\varphi} \big] d\gamma \\
& + \int_{\partial\Gamma_D(\boldsymbol{\phi}) \cup \Theta(\boldsymbol{\phi})} \{ (\mathbf{u} - \mathbf{u}_D) \cdot (\mathbf{S}(\mathbf{v}_0) \mathbf{v}) \\
& + (\mathbf{v}_0 - \mathbf{u}_D) \cdot (\mathbf{S}(\mathbf{u}) \mathbf{v}) \} \boldsymbol{\tau} \cdot \boldsymbol{\varphi} d\zeta
\end{aligned}$$

using Eq. (9.3.21), representing the result of Proposition 9.3.10, and Eq. (9.3.27) of Proposition 9.3.13. Here, we denote $(\mathbf{v} \cdot \nabla) \mathbf{u} = (\nabla \mathbf{u}^\top)^\top \mathbf{v}$ as $\partial_v \mathbf{u}$,

$$\begin{aligned}
\bar{w}(\boldsymbol{\varphi}, \mathbf{u}) &= -\mathbf{S}(\mathbf{u}) \left[\sum_{i \in \{1, \dots, d-1\}} \left\{ \boldsymbol{\tau}_i \cdot (\nabla \boldsymbol{\varphi}^\top \mathbf{v}) \right\} \boldsymbol{\tau}_i \right] \\
& + (\mathbf{v} \cdot \boldsymbol{\varphi}) (\nabla^\top \mathbf{S}(\mathbf{u}))^\top,
\end{aligned} \tag{9.12.48}$$

and $(\nabla \cdot \boldsymbol{\varphi})_\tau$ as Eq. (9.2.6).

With the above results in mind, assume that \mathbf{u} is a weak solution of Problem 9.12.1 and that the self-adjoint relationship (Eq. (9.12.16)) holds. In this case, we can write Eq. (9.12.48) as

$$\begin{aligned}
\tilde{f}'_0(\boldsymbol{\phi})[\boldsymbol{\varphi}] &= \mathcal{L}_{0\boldsymbol{\phi}^*}(\boldsymbol{\phi}, \mathbf{u}, \mathbf{v}_0)[\boldsymbol{\varphi}] = \langle \bar{\mathbf{g}}_0, \boldsymbol{\varphi} \rangle \\
& = \int_{\partial\Omega(\boldsymbol{\phi})} \bar{\mathbf{g}}_{\partial\Omega 0} \cdot \boldsymbol{\varphi} d\gamma + \int_{\Gamma_p(\boldsymbol{\phi})} \bar{\mathbf{g}}_{p0} \cdot \boldsymbol{\varphi} d\gamma \\
& + \int_{\partial\Gamma_p(\boldsymbol{\phi}) \cup \Theta(\boldsymbol{\phi})} \bar{\mathbf{g}}_{\partial p 0} \cdot \boldsymbol{\varphi} d\zeta + \int_{\Gamma_D(\boldsymbol{\phi})} \bar{\mathbf{g}}_{D0} \cdot \boldsymbol{\varphi} d\gamma
\end{aligned} \tag{9.12.49}$$

using the notation of Eq. (7.5.15) for \tilde{f}_0 and the Dirichlet condition in Problem 9.12.1, where

$$\bar{\mathbf{g}}_{\partial\Omega 0} = (-\mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{u}) + 2\mathbf{b} \cdot \mathbf{u}) \mathbf{v}, \tag{9.12.50}$$

$$\bar{\mathbf{g}}_{p0} = 2(\partial_v + \kappa) (\mathbf{p}_N \cdot \mathbf{u}) \mathbf{v}, \tag{9.12.51}$$

$$\bar{\mathbf{g}}_{\partial p 0} = 2(\mathbf{p}_N \cdot \mathbf{u}) \boldsymbol{\tau}, \tag{9.12.52}$$

$$\bar{\mathbf{g}}_{D0} = 2\{\partial_v(\mathbf{u} - \mathbf{u}_D) \cdot (\mathbf{S}(\mathbf{u}) \mathbf{v})\} \mathbf{v}. \tag{9.12.53}$$

Furthermore, on a homogeneous Dirichlet boundary, since there is a strain component only in the normal direction,

$$\partial_v \mathbf{u} = \mathbf{E}(\mathbf{u}) \mathbf{v} \tag{9.12.54}$$

holds. Hence, Eq. (9.12.53) can be written as

$$\bar{g}_{D0} = 2 \{ (\mathbf{E}(\mathbf{u}) \mathbf{v}) \cdot (\mathbf{S}(\mathbf{u}) \mathbf{v}) \} \mathbf{v} = 2 (\mathbf{E}(\mathbf{u}) \cdot \mathbf{S}(\mathbf{u})) \mathbf{v}. \quad (9.12.55)$$

Here, if \bar{g}_0 is written on the homogeneous Dirichlet boundary and homogeneous Neumann boundary, we get

$$\bar{g}_0 = (-\mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{u}) + 2\mathbf{b} \cdot \mathbf{u}) \mathbf{v} \quad \text{on } \Gamma_N(\phi) \setminus \bar{\Gamma}_p(\phi), \quad (9.12.56)$$

$$\bar{g}_0 = (\mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{u}) + 2\mathbf{b} \cdot \mathbf{u}) \mathbf{v} \quad \text{on } \Gamma_D(\phi). \quad (9.12.57)$$

From these results, it is evident that the sign of strain energy density $\mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{u}) / 2$ swaps between the homogeneous Dirichlet boundary and homogeneous Neumann boundary.

From the above results, conclusions similar to Theorem 9.8.6 can be obtained with respect to the function space containing \bar{g}_0 of Eq. (9.12.49).

On the other hand, the shape derivative of $f_1(\phi)$ can be written as

$$f'_1(\phi)[\varphi] = \langle \bar{g}_1, \varphi \rangle = \int_{\partial\Omega(\phi)} \bar{g}_{\partial\Omega 1} \cdot \varphi \, d\gamma, \quad (9.12.58)$$

where

$$\bar{g}_{\partial\Omega 1} = \mathbf{v}. \quad (9.12.59)$$

This can be obtained by letting $u = 1$ in Proposition 9.3.9 and is actually due to the fact that the solution for the state determination problem is not used.

9.12.4 Relation with Optimal Design Problem of Stepped One-Dimensional Linear Elastic Body

Let us think about the relationship between the shape derivative of the cost function in the mean compliance minimization problem (Problem 9.12.2) of a $d \in \{2, 3\}$ -dimensional linear elastic body and the cross-sectional derivative of the cost function in the mean compliance minimization problem (Problem 1.1.4) of the stepped one-dimensional linear elastic body seen in Chap. 1. Table 9.2 shows some comparisons between the two problems.

In Problem 1.1.4, the body force and known displacement were not used. If this assumption is applied to Problem 9.12.2, it corresponds to putting the cost function as

$$f_0(\phi, \mathbf{u}) = \hat{l}(\phi)(\mathbf{u}) = \int_{\Gamma_p(\phi)} \mathbf{p}_N \cdot \mathbf{u} \, d\gamma = \sum_{i \in \{1, 2\}} \int_{\Gamma_i} \frac{p_i}{a_i} u_i \, d\gamma, \quad (9.12.60)$$

Table 9.2 Correspondence between cross-sectional optimization problem and shape optimization problem

Comparison item	Cross-sectional optimization	Shape optimization
Design variable	$\mathbf{a} \in X = \mathbb{R}^2$	$\boldsymbol{\phi} \in X = H^1(D; \mathbb{R}^d)$
State variable	$\mathbf{u} \in U = \mathbb{R}^2$	$\mathbf{u} \in U = H^1(D; \mathbb{R}^d)$
State determination	$\mathcal{L}_S(\mathbf{a}, \mathbf{u}, \mathbf{v}) = 0 \forall \mathbf{v} \in U$	$\mathcal{L}_S(\boldsymbol{\phi}, \mathbf{u}, \mathbf{v}) = 0 \forall \mathbf{v} \in U$
Object function	$f_0 = \mathbf{p} \cdot \mathbf{u}$	$f_0 = \hat{l}(\boldsymbol{\phi})(\mathbf{u})$
Constraint function	$f_1 = (\text{volume}) - c_1$	$f_1 = (\text{domain measure}) - c_1$
Gradient	$\mathbf{g}_i \in X' = \mathbb{R}^2$	$\mathbf{g}_i, \tilde{\mathbf{g}}_i \in X' = H^{1'}(D; \mathbb{R}^d)$
Gradient method	$\mathbf{y}_{gi} \cdot \mathbf{A}z = -\mathbf{g} \cdot \mathbf{z} \forall \mathbf{z} \in X$	$a(\varphi_{gi}, \mathbf{z}) = -\langle \mathbf{g}_i, \mathbf{z} \rangle \forall \mathbf{z} \in X$

where p_i , u_i , a_i with respect to $i \in \{1, 2\}$ follow the respective definitions in Problem 1.1.4. Moreover, in Problem 1.1.4, external forces p_1 and p_2 were fixed with respect to the variation of the cross-sectional area. In other words, p_1 and p_2 are assumed to vary with boundary measures (Definition 9.4.4). In this case, the shape derivative of $f_{0\phi}(\boldsymbol{\phi}, \mathbf{u})[\varphi]$ becomes zero. On the other hand, \mathbf{p}_N was assumed to be fixed with the material (Definition 9.4.1) in Eq. (9.12.19) which gives the shape derivative of f_0 . Considering their differences, the shape derivative of f_0 defined by Eq. (9.12.60) can be written as

$$\tilde{f}'_0(\boldsymbol{\phi})[\varphi] = \langle \mathbf{g}_0, \varphi \rangle = \sum_{i \in \{1, 2\}} \int_0^l \left\{ \mathbf{G}_{\Omega 0i} \cdot (\nabla \varphi_i^\top) + g_{\Omega 0i} \nabla \cdot \varphi_i \right\} a_i \, dx. \quad (9.12.61)$$

Here, we assume that the cross-section of the stepped one-dimensional linear elastic body is a rectangle with unit depth. In addition, the x -coordinate is viewed as the x_1 -coordinate, and the height direction is viewed as the x_2 -coordinate. For each $i \in \{1, 2\}$, a_i represents the cross-section and b_i represents its variation. Moreover, $\sigma_1, \varepsilon_1, \sigma_2, \varepsilon_2$ denotes $\sigma(u_1), \varepsilon(u_1), \sigma(u_2 - u_1), \varepsilon(u_2 - u_1)$, respectively. In this case, the following relationships:

$$\mathbf{G}_{\Omega 0i} = 2\mathbf{S}(\mathbf{u}) \left(\nabla \mathbf{u}^\top \right)^\top = 2 \begin{pmatrix} \sigma_i & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \varepsilon_i & 0 \\ 0 & 0 \end{pmatrix} = 2 \begin{pmatrix} \sigma_i \varepsilon_i & 0 \\ 0 & 0 \end{pmatrix},$$

$$g_{\Omega 0i} = -\mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{u}) = - \begin{pmatrix} \sigma_i & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \varepsilon_i & 0 \\ 0 & 0 \end{pmatrix} = -\sigma_i \varepsilon_i,$$

$$\nabla \varphi_i^\top = \begin{pmatrix} 0 & 0 \\ 0 & b_i/a_i \end{pmatrix}, \quad \nabla \cdot \varphi_i = (\nabla \cdot \varphi_i)_\tau = \frac{b_i}{a_i}$$

hold. Using these relationships, we get

$$\tilde{f}'_0(\boldsymbol{\phi})[\varphi] = \langle \mathbf{g}_0, \varphi \rangle = l \begin{pmatrix} -\sigma_1 \varepsilon_1 \\ -\sigma_2 \varepsilon_2 \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \mathbf{g}_0 \cdot \mathbf{b}. \quad (9.12.62)$$

Here, \mathbf{g}_0 on the right-hand side of Eq. (9.12.62) matches the cross-sectional gradient of Eq. (1.1.28).

Moreover, the shape derivative of $f_1(\boldsymbol{\phi})$ becomes

$$\begin{aligned} f'_1(\boldsymbol{\phi})[\boldsymbol{\varphi}] &= \sum_{i \in \{1,2\}} \langle \mathbf{g}_1, \boldsymbol{\varphi} \rangle = \int_0^l (\nabla \cdot \boldsymbol{\varphi}_i) a_i \, dx \\ &= l \begin{pmatrix} 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \mathbf{g}_1 \cdot \mathbf{b}. \end{aligned} \quad (9.12.63)$$

\mathbf{g}_1 on the right-hand side of Eq. (9.12.63) matches the cross-sectional gradient of Eq. (1.1.17).

Furthermore, the Hessian matrix of f_0 defined by Eq. (9.12.60) can be obtained as follows. For each $j \in \{1, 2\}$, the following hold:

$$\begin{aligned} \nabla \boldsymbol{\varphi}_{ji}^\top &= \begin{pmatrix} 0 & 0 \\ 0 & \frac{b_{ji}}{a_i} \end{pmatrix}, \quad \nabla \cdot \boldsymbol{\varphi}_{ji} = \frac{b_{ji}}{a_i}, \\ \mathbf{E}(\mathbf{u}) &= \begin{pmatrix} \varepsilon_i & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{S}(\mathbf{u}) = \begin{pmatrix} \sigma_i & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

The shape derivative of the first term on the right-hand side of Eq. (9.12.61) is calculated by Eq. (9.12.37). Hence, we get

$$\begin{aligned} h_0(\boldsymbol{\phi}, \mathbf{u}, \mathbf{v}_0)[\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2] &= \sum_{i \in \{1,2\}} \int_0^l \left[2\mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{u}) (\nabla \cdot \boldsymbol{\varphi}_2) (\nabla \cdot \boldsymbol{\varphi}_1) \right. \\ &\quad + \left(\mathbf{S}(\mathbf{u}) (\nabla \mathbf{u}^\top)^\top \right) \cdot \left\{ \nabla \boldsymbol{\varphi}_2^\top \nabla \boldsymbol{\varphi}_1^\top + \nabla \boldsymbol{\varphi}_1^\top \nabla \boldsymbol{\varphi}_2^\top + \nabla \boldsymbol{\varphi}_2^\top (\nabla \boldsymbol{\varphi}_1^\top)^\top \right. \\ &\quad \left. \left. + \nabla \boldsymbol{\varphi}_1^\top (\nabla \boldsymbol{\varphi}_2^\top)^\top - 4 \nabla \boldsymbol{\varphi}_2^\top \nabla \cdot \boldsymbol{\varphi}_1 - 4 \nabla \boldsymbol{\varphi}_1^\top \nabla \cdot \boldsymbol{\varphi}_2 \right\} \right] a_i \, dx \\ &= \begin{pmatrix} b_{11} \\ b_{12} \end{pmatrix} \cdot \left(2l \begin{pmatrix} \frac{\sigma_1 \varepsilon_1}{a_1} & 0 \\ 0 & \frac{\sigma_2 \varepsilon_2}{a_2} \end{pmatrix} \begin{pmatrix} b_{21} \\ b_{22} \end{pmatrix} \right) = \mathbf{b}_1 \cdot (\mathbf{H}_0 \mathbf{b}_2). \end{aligned} \quad (9.12.64)$$

The \mathbf{H}_0 on the right-hand side of Eq. (9.12.64) matches the Hessian matrix of Eq. (1.1.29).

Based on these comparisons, the image of the minimum point of Problem 9.12.2 is thought to be as that depicted in Fig. 9.16 using Fig. 1.4 of Exercise 1.1.7.

Figures 9.17 and 9.18 show the images of the H^1 gradient method for obtaining the domain variations $\boldsymbol{\varphi}_{g0}$ and $\boldsymbol{\varphi}_{g1}$ that decreases \tilde{f}_0 and f_1 , respectively. Figure 9.19 shows the image of the Lagrange multiplier λ_1 such that the constraint

Fig. 9.16 The image of a minimizer ϕ of the mean compliance minimization problem (Problem 9.12.2)

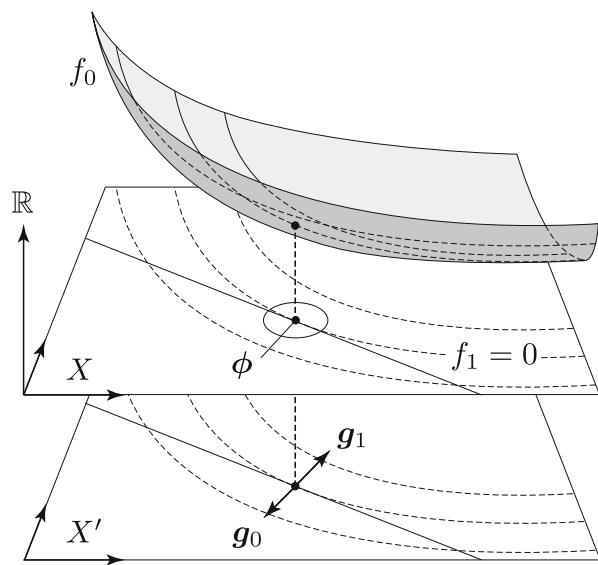
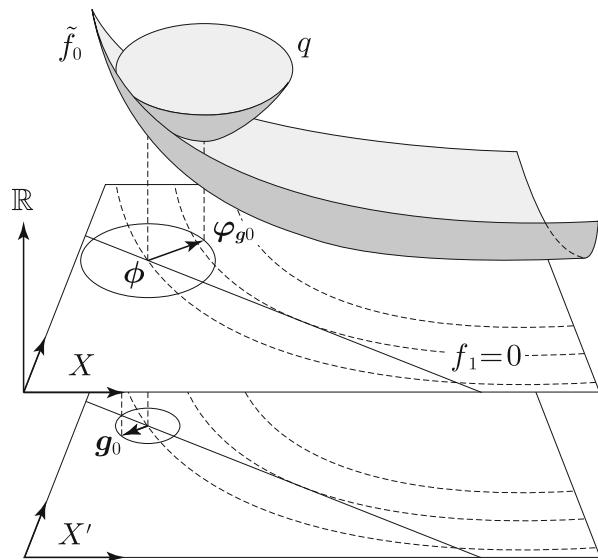


Fig. 9.17 The image of the H^1 gradient method with respect to \tilde{f}_0



concerning the domain measure is satisfied. In these figures, it is assumed that although the domain measure constraint is satisfied at $\Omega(\phi)$, ϕ is not a minimizer. The search direction $\varphi_g = \varphi_{g_0} + \lambda_1 \varphi_{g_1}$ in Fig. 9.19 is orthogonal to \mathbf{g}_1 in Fig. 9.18. In other words, the search direction is in the direction that the constraint is satisfied. This is due to the fact that Eq. (9.10.3) which determines the Lagrange multiplier in

Fig. 9.18 The image of the H^1 gradient method with respect to f_1

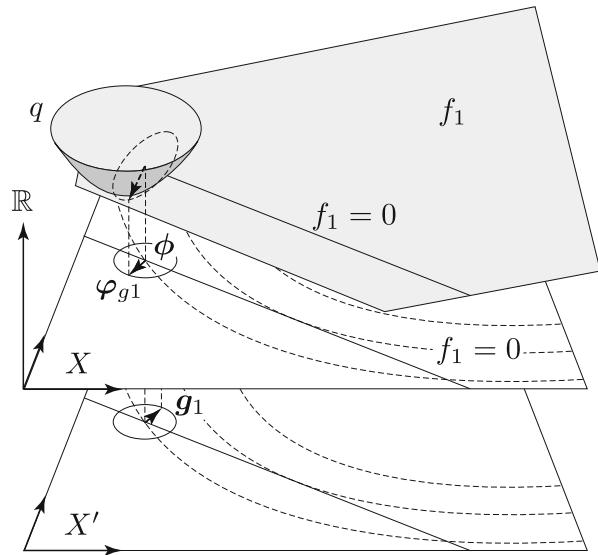
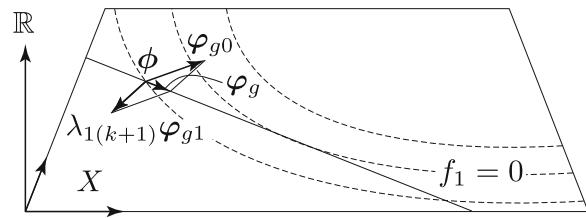


Fig. 9.19 Image of Lagrange multiplier λ_1



the gradient method with respect to a constrained problem is actually given by

$$\lambda_1 = -\frac{\langle \mathbf{g}_1, \boldsymbol{\varphi}_{g0} \rangle}{\langle \mathbf{g}_1, \boldsymbol{\varphi}_{g1} \rangle} \quad (9.12.65)$$

in Problem 9.12.2 and can be written as

$$\langle \mathbf{g}_1, \boldsymbol{\varphi}_{g0} + \lambda_{1(k+1)} \boldsymbol{\varphi}_{g1} \rangle = 0. \quad (9.12.66)$$

9.12.5 Numerical Example

Let us show a numerical example. In Figs. 9.20, 9.21, 9.22, the results of the mean compliance minimization with respect to a two-dimensional linear elastic body with a boundary condition referred to as the coat-hanging problem are shown.

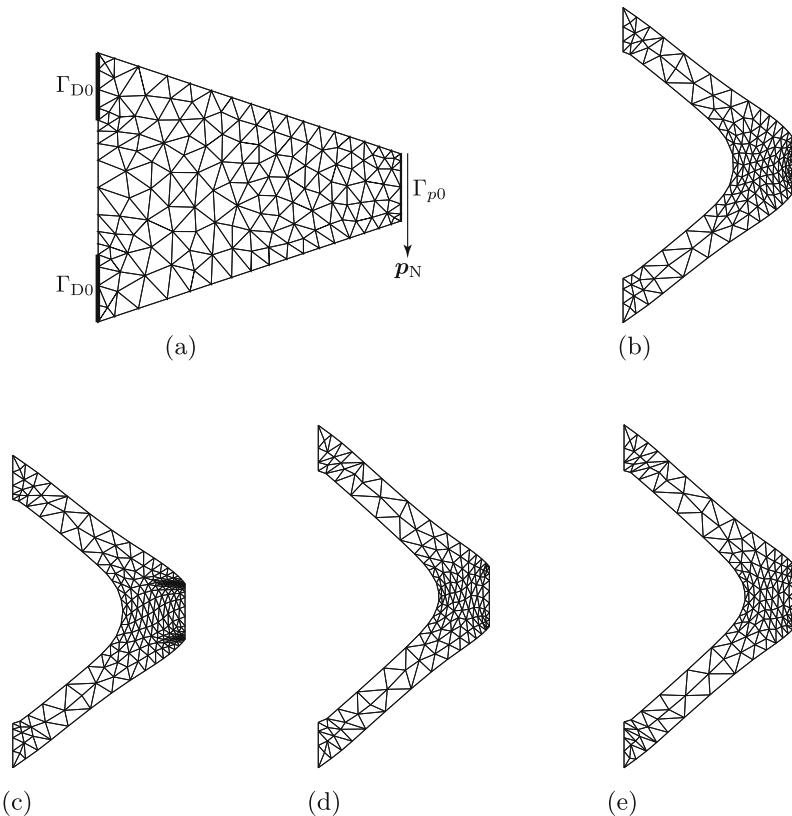


Fig. 9.20 Numerical example of mean compliance minimization problem: shape ($k = 200$). (a) Initial shape and boundary conditions. (b) H^1 gradient method ($\tilde{g}_\mathcal{L}$). (c) H^1 gradient method ($g_\mathcal{L}$). (d) H^1 Newton method ($h_0, g_\mathcal{L}$). (e) H^1 Newton method ($g_{h0}, g_\mathcal{L}$)

Figure 9.20a shows the initial shape and the boundary conditions of the state determination problem. The boundary condition with respect to the domain variation is assumed to be $\bar{\Omega}_{C0} = \Gamma_{D0} \cup \Gamma_{p0}$ in Eq. (9.1.1). Here, it is assumed that these boundaries deform in the tangential direction. In addition, p_N is assumed to vary with boundary measure. The program is written using the programming language FreeFEM (<https://freefem.org/>) [66] by the finite element method with reference to Example 37 in the book [130]. In the finite element analyses of the linear elastic problem and the H^1 gradient method or the H^1 Newton method, second-order triangular elements were used. In the case using the H^1 Newton method, the routine of the H^1 Newton method was started at $k_N = 120$. The parameters (c_a in Eq. (9.10.1), c_Ω in Eq. (9.9.3), k_N , $c_{\Omega 1}$ and $c_{\Omega 20}$ in Eq. (9.9.17), c_h in Eq. (9.10.8)

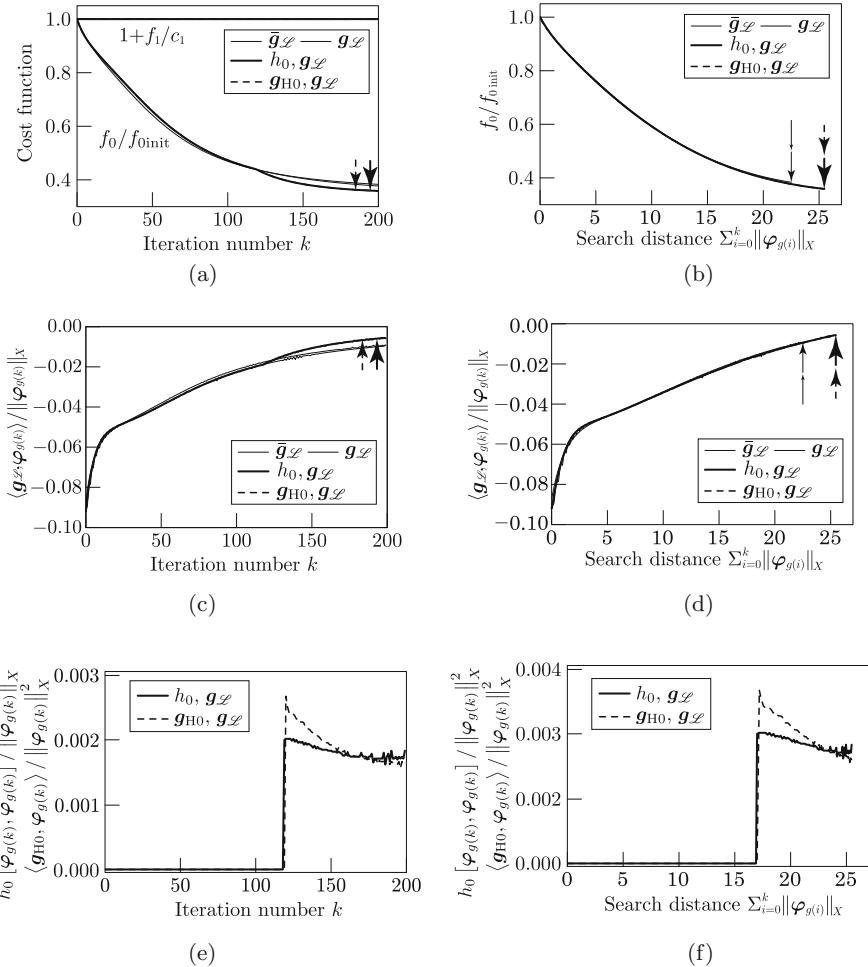


Fig. 9.21 Numerical example of mean compliance minimization problem: cost functions, their gradients and Hessians on the search path ($\bar{g}_{\mathcal{L}}$: H^1 gradient method using $\bar{g}_{\mathcal{L}}$, $g_{\mathcal{L}}$: H^1 gradient method using $g_{\mathcal{L}}$, $h_0, g_{\mathcal{L}}$: H^1 Newton method, $g_{H0}, g_{\mathcal{L}}$: H^1 Newton method using Hesse gradient). (a) Cost functions. (b) Cost functions (search distance). (c) Gradient of f_0 on search path (d) Gradient of f_0 on search path (search distance). (e) Hessian of f_0 on search path. (f) Hessian of f_0 on search path (search distance)

and the parameter (*errelas*) that controls the error level in the adaptive mesh) affect the result. The details are described in the programs.⁴

Figures 9.20b–e show the shapes obtained by the four methods (H^1 gradient method using $\bar{g}_{\mathcal{L}} = \bar{g}_0 + \lambda_1 \bar{g}_1$ of the boundary integral type, H^1 gradient method

⁴See Electronic supplementary material.

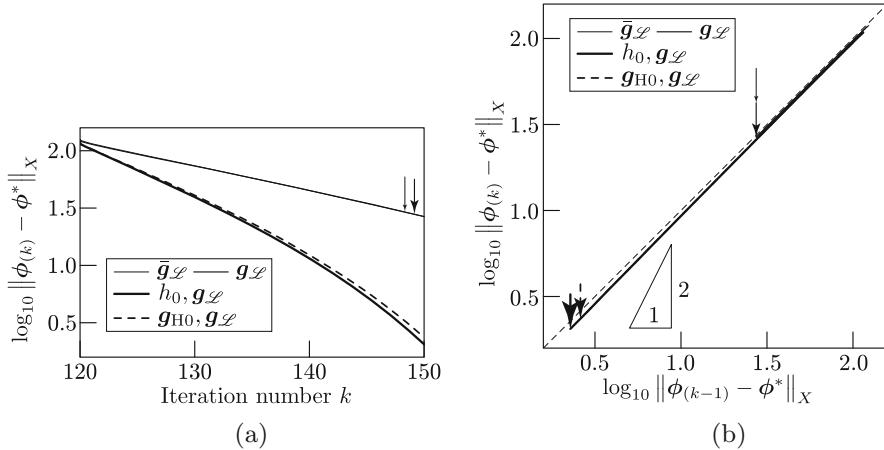


Fig. 9.22 Numerical example of mean compliance minimization problem: distance $\|\phi_{(k)} - \phi^*\|_X$ from an approximate minimum point ϕ^* ($\bar{g}_\mathcal{L}$: H^1 gradient method using $\bar{g}_\mathcal{L}$, $g_\mathcal{L}$: H^1 gradient method using $g_\mathcal{L}$, $h_0, g_\mathcal{L}$: H^1 Newton method, $g_{H0}, g_\mathcal{L}$: H^1 Newton method using Hesse gradient). (a) Iteration history. (b) $(k-1)$ -th vs. k -th plot

using $g_\mathcal{L} = g_0 + \lambda_1 g_1$ of the domain integral type, H^1 Newton method using $h_\mathcal{L} = h_0 + \lambda_1 h_1$ and $g_\mathcal{L}$, and H^1 Newton method using g_{H0}, h_1 and $g_\mathcal{L}$.

Figure 9.21a shows the cost functions $f_0/f_{0\text{init}}$ and $1 + f_1/c_1$ normalized with f_0 at the initial shape denoted by $f_{0\text{init}}$ and c_1 set with the initial volume, respectively, with respect to the iteration number k . Figure 9.21b shows those values with respect to the distance $\sum_{i=0}^{k-1} \|\varphi_{g(i)}\|_X$ on the search path in X . The graphs of f_0 's gradient (the gradient of the Lagrange function $\mathcal{L} = \mathcal{L}_0 + \lambda_1 f_1$) calculated as $\langle g_\mathcal{L}, \varphi_{g(k)} \rangle / \|\varphi_{g(k)}\|_X$ are shown in Fig. 9.21c,d with respect to the iteration number and the search distance, respectively. Moreover, Fig. 9.21e,f shows the graphs of f_0 's second-order derivative $h_0 [\varphi_{g(k)}, \varphi_{g(k)}] / \|\varphi_{g(k)}\|_X^2$ (in the case of the Newton method using Hesse gradient, $\langle g_{H0}, \varphi_{g(k)} \rangle / \|\varphi_{g(k)}\|_X^2$) with respect to the iteration number and the search distance, respectively. In these notations, the norm of the i -th search vector is defined by

$$\|\varphi_{g(i)}\|_X = \left(\int_{\Omega(\phi)} \left\{ \left(\nabla \varphi_{g(i)}^\top \right) \cdot \left(\nabla \varphi_{g(i)}^\top \right) + \varphi_{g(i)} \cdot \varphi_{g(i)} \right\} dx \right)^{1/2}. \quad (9.12.67)$$

The computational times until $k = 200$ by PC were 24.443, 37.132, 46.026, 59.312 sec when the H^1 gradient method of the boundary integral type, the H^1 gradient method of the domain integral type, the H^1 Newton method and the H^1 Newton method using the Hesse gradient were used, respectively.

Regarding the computational results obtained from the above-mentioned methods, we give the the following explanations and provide some considerations. The graphs in Fig. 9.21a show that the convergence speed with respect to the iteration

number k is faster when using the H^1 Newton method than when applying the H^1 gradient method. However, when the H^1 Newton method started, $c_{\Omega 1}$ and $c_{\Omega 0}$ in Eq. (9.9.17) were replaced with smaller values (the step size was enlarged) within the area where numerical instability did not happen. As a result, it can be considered that the convergence speed was increased. In this problem, when c_h is set to zero (that is, the H^1 gradient method), it was observed that the convergence speed was increased. The reason we consider to be behind the increase in convergence speed is the increase in the magnitude of the step size which was due to the exclusion of the term $h_{\mathcal{L}}$. In most cases, it is observed that when the step size is taken bigger, the H^1 Newton method keeps the computation until termination, but the H^1 gradient method fails to continue after a number of iterations. Moreover, the aspect around the minimum point can be observed in Fig. 9.21d,f. From these graphs, based on the observation that the Hessian of f_0 on the search path is positive valued, we infer that the point of convergence is a local minimum point.

In addition, Fig. 9.22a shows the graphs of the distance $\|\phi_{(k)} - \phi^*\|_X$ from the k -th approximation $\phi_{(k)}$ to an approximate minimum point ϕ^* obtained by the four methods with respect to the iteration number k . The approximate minimum point ϕ^* is given as the numerical solution of ϕ when the iteration time is taken larger than the given value in the H^1 Newton method. From this figure, it can be confirmed that the convergence orders for the results obtained through the H^1 Newton method are more than the first order. However, Fig. 9.22b, plotting the k -th distance $\|\phi_k - \phi^*\|_X$ with respect to the $(k - 1)$ -th distance (the gradient of the graph shows the order of convergence as explained by using Eq. (3.8.13)) shows that using Eq. (3.8.13)) shows that the convergence order of the H^1 Newton method is reason behind this finding is provided at the end of Sect. 8.9.6. Namely, the addition of the bilinear form a_X in X to the original Hessian in order to ensure coercivity and regularity of the left-hand side of Eq. (9.9.16) makes the H^1 Newton method different from the original Newton method.

9.13 Shape Optimization Problem of Stokes Flow Field

As an example of an application in flow field problems, let us consider a mean flow resistance minimization problem of a Stokes flow field and look at the process for obtaining the shape derivatives of cost functions. The image of the initial domain Ω_0 is shown in Fig. 9.23. The linear space X with respect to domain movement and its admissible set \mathcal{D} are defined as in Sect. 9.1.

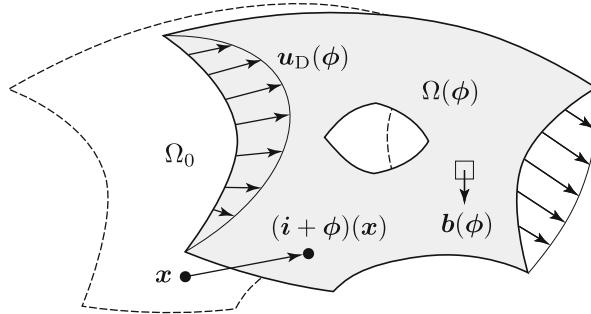


Fig. 9.23 The initial domain $\Omega_0 \subset D$ and the domain variation (displacement) ϕ with respect to a Stokes flow field

9.13.1 State Determination Problem

Let us consider a Stokes problem as a state determination problem. Here, in addition to the symbols used in Problem 5.5.1, the Stokes problem will be written in the following way for the shape optimization problem. Here, to guarantee the unique existence of the solution, $\partial\Omega(\phi)$ is taken to be a Dirichlet boundary with respect to $\phi \in \mathcal{D}$ and $\mathbf{u}_D : \partial\Omega(\phi) \rightarrow \mathbb{R}^d$ is taken to be a known flow velocity. Detailed conditions will be shown in Eqs. (9.13.5) and (9.13.6) later. μ is a positive constant expressing the coefficient of viscosity. With respect to the flow velocity \mathbf{u} , which is the solution to the state determination problem shown later, let $\mathbf{u} - \mathbf{u}_D$ be denoted as $\tilde{\mathbf{u}}$. Here, let the admissible set and the Hilbert space containing $\tilde{\mathbf{u}}$ be defined as

$$U = \left\{ \mathbf{u} \in H^1(D; \mathbb{R}^d) \mid \mathbf{u} = \mathbf{0}_{\mathbb{R}^d} \text{ on } \partial\Omega(\phi) \right\}, \quad (9.13.1)$$

$$\mathcal{S} = U \cap W^{2,4}(D; \mathbb{R}^d), \quad (9.13.2)$$

respectively. Moreover, the admissible set and the real Hilbert space containing the pressure p are taken to be

$$P = \left\{ q \in L^2(D; \mathbb{R}) \mid \int_{\Omega(\phi)} q \, dx = 0 \right\}, \quad (9.13.3)$$

$$\mathcal{Q} = P \cap W^{1,4}(D; \mathbb{R}), \quad (9.13.4)$$

respectively. For known functions, in conjunction with Hypothesis 9.5.1, it is assumed that

$$\mathbf{b} \in C_{S'}^1(B; L^\infty(D; \mathbb{R}^d)), \quad \mathbf{u}_D \in C_{S'}^1(B; U_{\text{div}} \cap C^{0,1}(D; \mathbb{R}^d)) \quad (9.13.5)$$

and these are fixed with the material. Moreover, with respect to Hypothesis 9.5.2,

$$\mathbf{b} \in C_{S^*}^1 \left(B; W^{1,2q_R} \left(D; \mathbb{R}^d \right) \right), \quad \mathbf{u}_D \in C_{S^*}^1 \left(B; U_{\text{div}} \cap W^{2,2q_R} \left(D; \mathbb{R}^d \right) \right) \quad (9.13.6)$$

and these are assumed to be fixed in space, where $q_R > d$. Here, let

$$U_{\text{div}} = \left\{ \mathbf{u} \in H^1 \left(D; \mathbb{R}^d \right) \mid \nabla \cdot \mathbf{u} = 0 \text{ in } D \right\}.$$

Here too, $(\mathbf{v} \cdot \nabla) \mathbf{u} = (\nabla \mathbf{u}^\top)^\top \mathbf{v}$ is written as $\partial_{\mathbf{v}} \mathbf{u}$. Given these definitions, we define a state determination problem as follows.

Problem 9.13.1 (Stokes Problem) For $\phi \in \mathcal{D}$, let \mathbf{b} , \mathbf{u}_D and μ be given. Find $(\mathbf{u}, p) : \Omega(\phi) \rightarrow \mathbb{R}^{d+1}$ which satisfies

$$\begin{aligned} -\nabla^\top \left(\mu \nabla \mathbf{u}^\top \right) + \nabla^\top p &= \mathbf{b}^\top(\phi) \quad \text{in } \Omega(\phi), \\ \nabla \cdot \mathbf{u} &= 0 \quad \text{in } \Omega(\phi), \\ \mathbf{u} &= \mathbf{u}_D(\phi) \quad \text{on } \partial \Omega(\phi), \\ \int_{\Omega(\phi)} p \, dx &= 0. \end{aligned} \quad \square$$

For later use, referring to the weak form of the Stokes problem (Problem 5.5.2) with a Dirichlet boundary condition, let the Lagrange function with respect to Problem 9.13.1 be

$$\begin{aligned} \mathcal{L}_S(\phi, \mathbf{u}, p, \mathbf{v}, q) &= \int_{\Omega(\phi)} \left\{ -\mu \nabla \mathbf{u}^\top \cdot (\nabla \mathbf{v}^\top) + p \nabla \cdot \mathbf{v} + q \nabla \cdot \mathbf{u} + \mathbf{b} \cdot \mathbf{v} \right\} dx \\ &\quad + \int_{\partial \Omega(\phi)} \{(\mathbf{u} - \mathbf{u}_D) \cdot (\mu \partial_{\mathbf{v}} \mathbf{v} - q \mathbf{v}) + \mathbf{v} \cdot (\mu \partial_{\mathbf{v}} \mathbf{u} - p \mathbf{v})\} d\gamma, \end{aligned} \quad (9.13.7)$$

where (\mathbf{u}, p) is not necessarily the solution of Problem 9.13.1 and (\mathbf{v}, q) is taken to be an element of $U \times P$ introduced as a Lagrange multiplier. If (\mathbf{u}, p) is the solution of Problem 9.13.1, the equation

$$\mathcal{L}_S(\phi, \mathbf{u}, p, \mathbf{v}, q) = 0$$

holds with respect to an arbitrary $(\mathbf{v}, q) \in U \times P$. This equation is equivalent to the weak form of Problem 9.13.1.

9.13.2 Mean Flow Resistance Minimization Problem

Let us define a shape optimization problem with the associated cost functions defined as follows. With respect to the solution (\mathbf{u}, p) of Problem 9.13.1,

$$f_0(\phi, \mathbf{u}, p) = - \int_{\Omega(\phi)} \mathbf{b} \cdot \mathbf{u} \, dx + \int_{\partial\Omega(\phi)} \mathbf{u}_D \cdot (\mu \partial_v \mathbf{u} - p \mathbf{v}) \, d\gamma \quad (9.13.8)$$

is referred to as the mean flow resistance. The reason for this is as explained in Sect. 8.10.2. Moreover,

$$f_1(\phi) = \int_{\Omega(\phi)} \, dx - c_1 \quad (9.13.9)$$

is a cost function with respect to domain measure constraint. Here, c_1 is a positive constant such that $f_1(\phi) \leq 0$ holds with respect to some $\phi \in \mathcal{D}$.

We define a mean flow resistance minimization problem as follows.

Problem 9.13.2 (Mean Flow Resistance Minimization Problem) Let \mathcal{D} , \mathcal{S} and \mathcal{Q} be defined as in Eq. (9.1.3), Eq. (9.13.2) and Eq. (9.13.4), respectively. Let f_0 and f_1 be Eq. (9.13.8) and Eq. (9.13.9), respectively. In this case, obtain $\Omega(\phi)$ which satisfies

$$\min_{(\phi, \mathbf{u} - \mathbf{u}_D, p) \in \mathcal{D} \times \mathcal{S} \times \mathcal{Q}} \{ f_0(\phi, \mathbf{u}, p) \mid f_1(\phi) \leq 0, \text{ Problem 9.13.1} \}. \quad \square$$

9.13.3 Shape Derivatives of Cost Functions

The shape derivative of $f_1(\phi)$ has already been obtained using Eq. (9.12.24) or Eq. (9.12.58). Hence, only the shape derivative of $f_0(\phi, \mathbf{u}, p)$ will be computed. Here too, let us consider the case of using the formulae based on the shape derivative of a function and the case using the formulae based on the partial shape derivative of a function separately. If the formulae based on the shape derivative of a function are used, the expression for the shape derivative up to the second order will be established. As preparation for this, let the Lagrange function of $f_0(\phi, \mathbf{u})$ be

$$\begin{aligned} \mathcal{L}_0(\phi, \mathbf{u}, p, \mathbf{v}_0, q_0) \\ = f_0(\phi, \mathbf{u}, p) - \mathcal{L}_S(\phi, \mathbf{u}, p, \mathbf{v}, q) \\ = \int_{\Omega(\phi)} \left\{ \mu \nabla \mathbf{u}^\top \cdot \nabla \mathbf{v}_0^\top - p \nabla \cdot \mathbf{v}_0 - \mathbf{b} \cdot (\mathbf{v}_0 + \mathbf{u}) - q_0 \nabla \cdot \mathbf{u} \right\} \, dx \\ - \int_{\partial\Omega(\phi)} \{ (\mathbf{u} - \mathbf{u}_D) \cdot (\mu \partial_v \mathbf{v}_0 - q_0 \mathbf{v}) + (\mathbf{v}_0 - \mathbf{u}_D) \cdot (\mu \partial_v \mathbf{u} - p \mathbf{v}) \} \, d\gamma. \end{aligned} \quad (9.13.10)$$

Here, \mathcal{L}_S is the Lagrange function of the state determination problem defined in Eq. (9.13.7). Moreover, it is assumed that (\mathbf{v}_0, q_0) is a Lagrange multiplier with respect to the state determination problem prepared for f_0 and that $(\mathbf{v}_0 - \mathbf{u}_D, q_0)$ is an element of $U \times P$.

Shape Derivative of f_0 Using Formulae Based on Shape Derivative of a Function

If the formulae based on the shape derivative of a function are used, the following results are obtained. In this case, it is assumed that \mathbf{b} and \mathbf{u}_D are fixed with the material.

In this case, the Fréchet derivative of \mathcal{L}_0 can be written as

$$\begin{aligned} \mathcal{L}'_0(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0) & [\boldsymbol{\varphi}, \hat{\mathbf{u}}, \hat{p}, \hat{\mathbf{v}}_0, \hat{q}_0] \\ &= \mathcal{L}_{0\boldsymbol{\varphi}'}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0) [\boldsymbol{\varphi}] + \mathcal{L}_{0\mathbf{u}p}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0) [\hat{\mathbf{u}}, \hat{p}] \\ &\quad + \mathcal{L}_{0\mathbf{v}_0q_0}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0) [\hat{\mathbf{v}}_0, \hat{q}_0] \end{aligned} \quad (9.13.11)$$

for any arbitrary variation $(\boldsymbol{\varphi}, \hat{\mathbf{u}}, \hat{p}, \hat{\mathbf{v}}_0, \hat{q}_0) \in X \times (U \times P)^2$. Here, the notations in Eqs. (9.3.5) and (9.3.15) were used. Each term is considered below.

The third term on the right-hand side of Eq. (9.13.11) becomes

$$\begin{aligned} \mathcal{L}_{0\mathbf{v}_0q_0}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0) [\hat{\mathbf{v}}_0, \hat{q}_0] &= -\mathcal{L}_{S\mathbf{v}_0,q_0}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0) [\hat{\mathbf{v}}_0, \hat{q}_0] \\ &= -\mathcal{L}_S(\boldsymbol{\phi}, \mathbf{u}, p, \hat{\mathbf{v}}_0, \hat{q}_0). \end{aligned} \quad (9.13.12)$$

Equation (9.13.12) is the Lagrange function of the state determination problem (Problem 9.13.1). Hence, if (\mathbf{u}, p) is the weak solution of the state determination problem, the third term on the right-hand side of Eq. (9.13.11) is zero.

Moreover, the second term on the right-hand side of Eq. (9.13.11) becomes

$$\begin{aligned} \mathcal{L}_{0\mathbf{u}p}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0) [\hat{\mathbf{u}}, \hat{p}] &= \int_{\Omega(\boldsymbol{\phi})} \left\{ \mu \left(\nabla \mathbf{u}'^\top \right) \cdot \nabla \mathbf{v}_0^\top - \hat{p} \nabla \cdot \mathbf{v}_0 - \mathbf{b} \cdot \hat{\mathbf{u}} - q_0 \nabla \cdot \hat{\mathbf{u}} \right\} dx \\ &\quad - \int_{\partial\Omega(\boldsymbol{\phi})} \{ \hat{\mathbf{u}} \cdot (\mu \partial_\nu \mathbf{v}_0 - q_0 \mathbf{v}) + (\mathbf{v}_0 - \mathbf{u}_D) \cdot (\mu \partial_\nu \hat{\mathbf{u}} - \hat{p} \mathbf{v}) \} d\gamma \\ &= -\mathcal{L}_S(\boldsymbol{\phi}, \mathbf{v}_0, q_0, \hat{\mathbf{u}}, \hat{p}) \end{aligned} \quad (9.13.13)$$

for any arbitrary variation $(\hat{\mathbf{u}}, \hat{p}) \in U \times P$ of (\mathbf{u}, p) . Hence, when the self-adjoint relationship

$$(\mathbf{u}, p) = (\mathbf{v}_0, q_0) \quad (9.13.14)$$

holds, the second term on the right-hand side of Eq. (9.13.11) also vanishes.

Furthermore, the first term on the right-hand side of Eq. (9.13.11) becomes

$$\begin{aligned} & \mathcal{L}_{0\phi'}(\phi, \mathbf{u}, p, \mathbf{v}_0, q_0)[\phi] \\ &= \int_{\Omega(\phi)} \left[-\mu \nabla \mathbf{u}^\top \cdot (\nabla \phi^\top \nabla \mathbf{v}_0^\top) - \mu \nabla \mathbf{v}_0^\top \cdot (\nabla \phi^\top \nabla \mathbf{u}^\top) \right. \\ & \quad + p (\nabla \phi^\top \nabla) \cdot \mathbf{v}_0 + q_0 (\nabla \phi^\top \nabla) \cdot \mathbf{u} \\ & \quad + \left\{ \mu \nabla \mathbf{u}^\top \cdot \nabla \mathbf{v}_0^\top - p \nabla \cdot \mathbf{v}_0 - q_0 \nabla \cdot \mathbf{u} - \mathbf{b} \cdot (\mathbf{u} + \mathbf{v}_0) \right\} \nabla \cdot \phi \] dx \\ & \quad - \int_{\partial\Omega(\phi)} \left[\{(\mathbf{u} - \mathbf{u}_D) \cdot \mathbf{w}(\phi, \mathbf{v}_0, q_0) + (\mathbf{v}_0 - \mathbf{u}_D) \cdot \mathbf{w}(\phi, \mathbf{u}, p)\} \right. \\ & \quad + \{(\mathbf{u} - \mathbf{u}_D) \cdot (\mu \partial_\nu \mathbf{v}_0 - q_0 \mathbf{v}_0) \\ & \quad \left. + (\mathbf{v}_0 - \mathbf{u}_D) \cdot (\mu \partial_\nu \mathbf{u} - p \mathbf{v}) \right\} (\nabla \cdot \phi)_\tau \] d\gamma, \end{aligned}$$

in view of Eqs. (9.3.5) and (9.3.15) representing the results of Propositions 9.3.4 and 9.3.7. Here, we let

$$\begin{aligned} & \mathbf{w}(\phi, \mathbf{u}, p) \\ &= \left\{ (\mu \nabla \mathbf{u}^\top)^\top - p \mathbf{I} \right\} \left[\left\{ \mathbf{v} \cdot (\nabla \phi^\top \mathbf{v}) \right\} \mathbf{v} - \left\{ \nabla \phi^\top + (\nabla \phi^\top)^\top \right\} \mathbf{v} \right], \end{aligned} \quad (9.13.15)$$

and $(\nabla \cdot \phi)_\tau$ follows Eq. (9.2.6). \mathbf{I} represents a d -order unit matrix. Furthermore, by applying the identity

$$(\nabla \phi^\top \nabla) \cdot \mathbf{v}_0 = (\nabla \mathbf{v}_0^\top)^\top \cdot \nabla \phi^\top = \mathbf{I} \cdot (\nabla \phi^\top \nabla \mathbf{v}_0^\top), \quad (9.13.16)$$

we get

$$\begin{aligned} & \mathcal{L}_{0\phi'}(\phi, \mathbf{u}, p, \mathbf{v}_0, q_0)[\phi] \\ &= \int_{\Omega(\phi)} \left[-(\mu \nabla \mathbf{u}^\top - p \mathbf{I}) \cdot (\nabla \phi^\top \nabla \mathbf{v}_0^\top) - (\mu \nabla \mathbf{v}_0^\top - q_0 \mathbf{I}) \cdot (\nabla \phi^\top \nabla \mathbf{u}^\top) \right. \\ & \quad + \left\{ (\mu \nabla \mathbf{u}^\top - p \mathbf{I}) \cdot \nabla \mathbf{v}_0^\top - q_0 \nabla \cdot \mathbf{u} - \mathbf{b} \cdot (\mathbf{u} + \mathbf{v}_0) \right\} \nabla \cdot \phi \] dx \end{aligned}$$

$$\begin{aligned}
& - \int_{\partial\Omega(\phi)} \left[\{(\mathbf{u} - \mathbf{u}_D) \cdot \mathbf{w}(\phi, \mathbf{v}_0, q_0) + (\mathbf{v}_0 - \mathbf{u}_D) \cdot \mathbf{w}(\phi, \mathbf{u}, p)\} \right. \\
& + \{(\mathbf{u} - \mathbf{u}_D) \cdot (\mu \partial_v \mathbf{v}_0 - q_0 \mathbf{v}) \\
& \left. + (\mathbf{v}_0 - \mathbf{u}_D) \cdot (\mu \partial_v \mathbf{u} - p \mathbf{v})\} (\nabla \cdot \phi)_\tau \right] d\gamma. \tag{9.13.17}
\end{aligned}$$

With the above results in mind, it is assumed that (\mathbf{u}, p) is the weak solution of Problem 9.13.1 and that the self-adjoint relationship (Eq. (9.13.14)) holds true. Here, using the Dirichlet condition and the continuity equation of Problem 9.13.1, we get

$$\begin{aligned}
\tilde{f}'_0(\phi)[\phi] &= \mathcal{L}_{0\phi'}(\phi, \mathbf{u}, p, \mathbf{v}_0, q_0)[\phi] = \langle \mathbf{g}_0, \phi \rangle \\
&= \int_{\Omega(\phi)} \left(\mathbf{G}_{\Omega 0} \cdot \nabla \phi^\top + g_{\Omega 0} \nabla \cdot \phi \right) dx \tag{9.13.18}
\end{aligned}$$

following the notation of Eq. (7.5.15) for \tilde{f}_0 , where

$$\mathbf{G}_{\Omega 0} = -2 \left(\mu \nabla \mathbf{u}^\top - p \mathbf{I} \right) \left(\nabla \mathbf{u}^\top \right)^\top, \tag{9.13.19}$$

$$g_{\Omega 0} = \mu \nabla \mathbf{u}^\top \cdot \nabla \mathbf{u}^\top - 2 \mathbf{b} \cdot \mathbf{u}. \tag{9.13.20}$$

From the above results, similar conclusions can be obtained for Theorem 9.8.2 with respect to \mathbf{g}_0 of Eq. (9.13.18).

Second-Order Shape Derivative of f_0 Using Formulae Based on Shape Derivative of a Function

Next, let us obtain the second-order shape derivative of mean flow resistance f_0 . Here, the formulae based on the shape derivative of a function are used, following the procedures shown in Sect. 9.8.2.

In correspondence with Hypothesis 9.8.3, the first condition $\mathbf{b} = \mathbf{0}_{\mathbb{R}^d}$ is assumed and we also suppose that the second condition is again satisfied. However, Hypothesis 9.8.3 (3) is unnecessary.

The Lagrange function \mathcal{L}_0 of f_0 is defined by Eq. (9.13.10). Viewing (ϕ, \mathbf{u}, p) as a design variable, we define its admissible set and admissible direction set respectively as

$$S = \{(\phi, \mathbf{u}, p) \in \mathcal{D} \times \mathcal{S} \times \mathcal{Q} \mid \mathcal{L}_S(\phi, \mathbf{u}, p, \mathbf{v}, q) = 0 \text{ for all } (\mathbf{v}, q) \in U \times P\},$$

$$T_S(\phi, \mathbf{u}, p) = \{(\phi, \hat{\mathbf{v}}, \hat{p}) \in X \times U \times P \mid$$

$$\mathcal{L}_{S\phi u p}(\phi, \mathbf{u}, p, \mathbf{v}, q) [\phi, \hat{\mathbf{v}}, \hat{p}] = 0 \text{ for all } (\mathbf{v}, q) \in U \times P\}.$$

Considering Eq. (9.1.6), the second-order Fréchet partial derivative of \mathcal{L}_0 with respect to arbitrary variations $(\boldsymbol{\varphi}_1, \hat{\mathbf{v}}_1, \hat{\pi}_1), (\boldsymbol{\varphi}_2, \hat{\mathbf{v}}_2, \hat{\pi}_2) \in T_S(\boldsymbol{\phi}, \mathbf{u}, p)$ of the design variable $(\boldsymbol{\phi}, \mathbf{u}, p) \in S$ is given as follows:

$$\begin{aligned}
& \mathcal{L}_{0(\boldsymbol{\phi}', \mathbf{u}, p)(\boldsymbol{\phi}', \mathbf{u}, p)}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0) [(\boldsymbol{\varphi}_1, \hat{\mathbf{v}}_1, \hat{\pi}_1), (\boldsymbol{\varphi}_2, \hat{\mathbf{v}}_2, \hat{\pi}_2)] \\
&= (\mathcal{L}_{0(\boldsymbol{\phi}', \mathbf{u}, p)}_{(\boldsymbol{\phi}', \mathbf{u}, p)}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0) [(\boldsymbol{\varphi}_1, \hat{\mathbf{v}}_1, \hat{\pi}_1), (\boldsymbol{\varphi}_2, \hat{\mathbf{v}}_2, \hat{\pi}_2)] \\
&\quad + \langle \mathbf{g}(\boldsymbol{\phi}), \mathbf{t}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) \rangle) \\
&= (\mathcal{L}_{0\boldsymbol{\phi}'}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0) [\boldsymbol{\varphi}_1] + \mathcal{L}_{0\mathbf{u}p}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0) [\hat{\mathbf{v}}_1, \hat{\pi}_1])_{\boldsymbol{\phi}'} [\boldsymbol{\varphi}_2] \\
&\quad + (\mathcal{L}_{0\boldsymbol{\phi}'}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0) [\boldsymbol{\varphi}_1] + \mathcal{L}_{0\mathbf{u}p}(\boldsymbol{\phi}, \mathbf{u}, \mathbf{v}_0) [\hat{\mathbf{v}}_1, \hat{\pi}_1])_{\mathbf{u}p} [\hat{\mathbf{v}}_2, \hat{\pi}_2] \\
&\quad + \langle \mathbf{g}(\boldsymbol{\phi}), \mathbf{t}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) \rangle \\
&= (\mathcal{L}_{0\boldsymbol{\phi}'}_{\boldsymbol{\phi}'}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0) [\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2] + \mathcal{L}_{0\boldsymbol{\phi}'\mathbf{u}p}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0) [\boldsymbol{\varphi}_1, \hat{\mathbf{v}}_2, \hat{\pi}_2] \\
&\quad + \mathcal{L}_{0\boldsymbol{\phi}'\mathbf{u}p}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0) [\boldsymbol{\varphi}_2, \hat{\mathbf{v}}_1, \hat{\pi}_1] \\
&\quad + \mathcal{L}_{0\mathbf{u}p\mathbf{u}p}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0) [\hat{\mathbf{v}}_1, \hat{\pi}_1, \hat{\mathbf{v}}_2, \hat{\pi}_2] \\
&\quad + \langle \mathbf{g}(\boldsymbol{\phi}), \mathbf{t}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) \rangle), \tag{9.13.21}
\end{aligned}$$

where $\langle \mathbf{g}_0(\boldsymbol{\phi}), \mathbf{t}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) \rangle$ follows the definition given in Eq. (9.1.8).

Here, the first and fifth terms of the right-hand side in Eq. (9.13.21) become

$$\begin{aligned}
& (\mathcal{L}_{0\boldsymbol{\phi}'}_{\boldsymbol{\phi}'}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0) [\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2] + \langle \mathbf{g}(\boldsymbol{\phi}), \mathbf{t}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) \rangle) \\
&= \int_{\Omega(\boldsymbol{\phi})} \left[\left\{ -\left(\mu \nabla \mathbf{u}^\top - p \mathbf{I} \right) \cdot \left(\nabla \boldsymbol{\varphi}_1^\top \nabla \mathbf{v}_0^\top \right) \right\}_{\boldsymbol{\phi}'} [\boldsymbol{\varphi}_2] \right. \\
&\quad + \left\{ -\left(\mu \nabla \mathbf{v}_0^\top - q_0 \mathbf{I} \right) \cdot \left(\nabla \boldsymbol{\varphi}_1^\top \nabla \mathbf{u}^\top \right) \right\}_{\boldsymbol{\phi}'} [\boldsymbol{\varphi}_2] \\
&\quad + \mu \nabla \mathbf{u}^\top \cdot \nabla \mathbf{v}_0^\top (\nabla \cdot \boldsymbol{\varphi}_1)_{\boldsymbol{\phi}'} [\boldsymbol{\varphi}_2] \\
&\quad - 2 \left(\mu \nabla \mathbf{u}^\top - p \mathbf{I} \right) \left(\nabla \mathbf{u}^\top \right)^\top \cdot \left(\nabla \boldsymbol{\varphi}_2^\top \nabla \boldsymbol{\varphi}_1^\top - \nabla \boldsymbol{\varphi}_1^\top (\nabla \cdot \boldsymbol{\varphi}_2) \right) \\
&\quad \left. + \mu \nabla \mathbf{u}^\top \cdot \nabla \mathbf{u}^\top \left\{ \left(\nabla \boldsymbol{\varphi}_2^\top \right)^\top \cdot \nabla \boldsymbol{\varphi}_1^\top - (\nabla \cdot \boldsymbol{\varphi}_2) (\nabla \cdot \boldsymbol{\varphi}_1) \right\} \right] dx, \tag{9.13.22}
\end{aligned}$$

which is obtained from Eq. (9.13.17) using the Dirichlet condition of Problem 9.13.1, the equation of continuity and the assumption $\mathbf{b} = \mathbf{0}_{\mathbb{R}^d}$. The first term of the integrand on the right-hand side of Eq. (9.13.22) can be expanded as follows:

$$\begin{aligned}
& \left\{ - \left(\mu \nabla \mathbf{u}^\top - p \mathbf{I} \right) \cdot \left(\nabla \boldsymbol{\varphi}_1^\top \nabla \mathbf{v}_0^\top \right) \right\}_{\boldsymbol{\varphi}'} [\boldsymbol{\varphi}_2] \\
&= \left\{ \left(\mu \nabla \mathbf{u}^\top - p \mathbf{I} \right) \cdot \left(\nabla \boldsymbol{\varphi}_1^\top \nabla \mathbf{v}_0^\top \right) \right\}_{\nabla \mathbf{u}^\top} \cdot \left(\nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{u}^\top \right) \\
&\quad + \left\{ \left(\mu \nabla \mathbf{u}^\top - p \mathbf{I} \right) \cdot \left(\nabla \boldsymbol{\varphi}_1^\top \nabla \mathbf{v}_0^\top \right) \right\}_{\nabla \boldsymbol{\varphi}_1^\top} \cdot \left(\nabla \boldsymbol{\varphi}_2^\top \nabla \boldsymbol{\varphi}_1^\top \right) \\
&\quad + \left\{ \left(\mu \nabla \mathbf{u}^\top - p \mathbf{I} \right) \cdot \left(\nabla \boldsymbol{\varphi}_1^\top \nabla \mathbf{v}_0^\top \right) \right\}_{\nabla \mathbf{v}_0^\top} \cdot \left(\nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{v}_0^\top \right) \\
&\quad - \left\{ \left(\mu \nabla \mathbf{u}^\top - p \mathbf{I} \right) \cdot \left(\nabla \boldsymbol{\varphi}_1^\top \nabla \mathbf{v}_0^\top \right) \right\} \nabla \cdot \boldsymbol{\varphi}_2 \\
&= \mu \left(\nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{u}^\top \right) \cdot \left(\nabla \boldsymbol{\varphi}_1^\top \nabla \mathbf{v}_0^\top \right) \\
&\quad + \left(\mu \nabla \mathbf{u}^\top - p \mathbf{I} \right) \cdot \left\{ \left(\nabla \boldsymbol{\varphi}_2^\top \nabla \boldsymbol{\varphi}_1^\top \nabla \mathbf{v}_0^\top \right) + \left(\nabla \boldsymbol{\varphi}_1^\top \nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{v}_0^\top \right) \right\} \\
&\quad - \left\{ \left(\mu \nabla \mathbf{u}^\top - p \mathbf{I} \right) \cdot \left(\nabla \boldsymbol{\varphi}_1^\top \nabla \mathbf{v}_0^\top \right) \right\} \nabla \cdot \boldsymbol{\varphi}_2. \tag{9.13.23}
\end{aligned}$$

Similarly, the second term of the integrand on the right-hand side of Eq. (9.13.22) is the first term with (\mathbf{u}, p) and (\mathbf{v}_0, q_0) switched over. The third term of the integrand on the right-hand side of Eq. (9.13.22) is

$$\begin{aligned}
& \mu \nabla \mathbf{u}^\top \cdot \nabla \mathbf{v}_0^\top \left\{ \nabla \cdot \boldsymbol{\varphi}_1 \right\}_{\boldsymbol{\varphi}'} [\boldsymbol{\varphi}_2] \\
&= \mu \nabla \mathbf{u}^\top \cdot \nabla \mathbf{v}_0^\top \left\{ - \left(\nabla \boldsymbol{\varphi}_2^\top \right)^\top \cdot \nabla \boldsymbol{\varphi}_1^\top + \left(\nabla \cdot \boldsymbol{\varphi}_2 \right) \left(\nabla \cdot \boldsymbol{\varphi}_1 \right) \right\}. \tag{9.13.24}
\end{aligned}$$

Hence, using the self-adjoint relationship, Eq. (9.13.22) becomes

$$\begin{aligned}
& \left(\mathcal{L}_{0\boldsymbol{\varphi}'} \right)_{\boldsymbol{\varphi}'} (\boldsymbol{\varphi}, \mathbf{u}, p, \mathbf{v}_0, q_0) [\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2] + \langle \mathbf{g}(\boldsymbol{\varphi}), \mathbf{t}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) \rangle \\
&= \int_{\Omega(\boldsymbol{\varphi})} \left[\mu \left(\nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{u}^\top \right) \cdot \left(\nabla \boldsymbol{\varphi}_1^\top \nabla \mathbf{v}_0^\top \right) \right. \\
&\quad + \left(\mu \nabla \mathbf{u}^\top - p \mathbf{I} \right) \cdot \left(\nabla \boldsymbol{\varphi}_1^\top \nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{v}_0^\top \right) \\
&\quad + \mu \left(\nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{v}_0^\top \right) \cdot \left(\nabla \boldsymbol{\varphi}_1^\top \nabla \mathbf{u}^\top \right) \\
&\quad \left. + \left(\mu \nabla \mathbf{v}_0^\top - q_0 \mathbf{I} \right) \cdot \left(\nabla \boldsymbol{\varphi}_1^\top \nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{u}^\top \right) \right] \mathrm{d}x. \tag{9.13.25}
\end{aligned}$$

Next, we consider the second term on the right-hand side of Eq. (9.13.21). Using Eq. (9.13.17), the Dirichlet condition of Problem 9.13.1, the equation of continuity and the assumption $\mathbf{b} = \mathbf{0}_{\mathbb{R}^d}$, we get

$$\begin{aligned} & \mathcal{L}_{0\phi'up}(\phi, \mathbf{u}, p, \mathbf{v}_0, q_0) [\varphi_1, \hat{\mathbf{v}}_2, \hat{\pi}_2] \\ &= \int_{\Omega(\phi)} \left[-\left(\mu \nabla \hat{\mathbf{v}}_2^\top - \hat{\pi}_2 \mathbf{I} \right) \cdot \left(\nabla \varphi_1^\top \nabla \mathbf{v}_0^\top \right) \right. \\ &\quad \left. - \left(\mu \nabla \mathbf{v}_0^\top - q_0 \mathbf{I} \right) \cdot \left(\nabla \varphi_1^\top \nabla \hat{\mathbf{v}}_2^\top \right) \right. \\ &\quad \left. + \left\{ \mu \left(\nabla \hat{\mathbf{v}}_2^\top - \hat{\pi}_2 \mathbf{I} \right) \cdot \nabla \mathbf{v}_0^\top - q_0 \nabla \cdot \hat{\mathbf{v}}_2 \right\} \nabla \cdot \varphi_1 \right] dx. \end{aligned} \quad (9.13.26)$$

On the other hand, the variation of (\mathbf{u}, p) satisfying the state determination problem with respect to an arbitrary domain variation $\varphi_j \in Y$ for $j \in \{1, 2\}$ is written as $(\hat{\mathbf{v}}_j, \hat{\pi}_j) = (\mathbf{v}'(\phi)[\varphi_j], \pi'(\phi)[\varphi_j])$. If the Fréchet partial derivative of the Lagrange function \mathcal{L}_S of the state determination problem defined in Eq. (9.13.7) is taken, then we obtain

$$\begin{aligned} & \mathcal{L}_{S\phi'up}(\phi, \mathbf{u}, p, \mathbf{v}, q) [\varphi_j, \hat{\mathbf{v}}_j, \hat{\pi}_j] \\ &= \int_{\Omega(\phi)} \left[\mu \left(\nabla \varphi_j^\top \nabla \mathbf{u}^\top \right) \cdot \left(\nabla \mathbf{v}^\top \right) + \mu \nabla \mathbf{u}^\top \cdot \left(\nabla \varphi_j^\top \nabla \mathbf{v}^\top \right) \right. \\ &\quad \left. - p \left(\nabla \varphi_j^\top \nabla \right) \cdot \mathbf{v} - q \left(\nabla \varphi_j^\top \nabla \right) \cdot \mathbf{u} \right. \\ &\quad \left. + \left\{ -\nabla \mathbf{u}^\top \cdot \left(\mu \nabla \mathbf{v}^\top - q \mathbf{I} \right) + p \nabla \cdot \mathbf{v} \right\} \nabla \cdot \varphi_j \right. \\ &\quad \left. - \left(\nabla \hat{\mathbf{v}}_j^\top \right) \cdot \left(\mu \nabla \mathbf{v}^\top - q \mathbf{I} \right) + \hat{\pi}_j \nabla \cdot \mathbf{v} \right] dx \\ &= \int_{\Omega(\phi)} \left[\left\{ \mu \left(\nabla \varphi_j^\top + \left(\nabla \varphi_j^\top \right)^\top - \left(\nabla \cdot \varphi_j \right) \mathbf{I} \right) \nabla \mathbf{u}^\top - \mu \nabla \hat{\mathbf{v}}_j^\top \right. \right. \\ &\quad \left. \left. + \hat{\pi}_j \mathbf{I} + p \left(\nabla \cdot \varphi_j \right) \mathbf{I} - p \left(\nabla \varphi_j^\top \right)^\top \right\} \cdot \nabla \mathbf{v}^\top \right. \\ &\quad \left. + q \left\{ - \left(\nabla \varphi_j^\top \nabla \right) \cdot \mathbf{u} + \left(\nabla \cdot \mathbf{u} \right) \left(\nabla \cdot \varphi_j \right) + \nabla \cdot \hat{\mathbf{v}}_j \right\} \right] dx \\ &= 0 \end{aligned} \quad (9.13.27)$$

for any arbitrary variation $(\mathbf{v}, q) \in U \times P$. From Eq. (9.13.27), the following identities:

$$\nabla \hat{\mathbf{v}}_j^\top = \left\{ \nabla \varphi_j^\top + \left(\nabla \varphi_j^\top \right)^\top - \nabla \cdot \varphi_j \right\} \nabla \mathbf{u}^\top, \quad (9.13.28)$$

$$\nabla \cdot \hat{\mathbf{v}}_j = \left(\nabla \varphi_j^\top \nabla \right) \cdot \mathbf{u} - (\nabla \cdot \mathbf{u}) (\nabla \cdot \varphi_j), \quad (9.13.29)$$

$$\hat{\pi}_j \mathbf{I} = -p (\nabla \cdot \varphi_j) \mathbf{I} - p \left(\nabla \varphi_j^\top \right)^\top \quad (9.13.30)$$

hold for any $(\mathbf{v}, q) \in U \times P$. Here, if $\hat{\mathbf{v}}_2$ and $\hat{\pi}_2$ satisfying Eq. (9.13.28) to Eq. (9.13.30) are substituted into $\hat{\mathbf{v}}_2$ and $\hat{\pi}_2$ of Eq. (9.13.26), we have

$$\begin{aligned} & \mathcal{L}_{0\phi' \mathbf{u} p} (\phi, \mathbf{u}, p, \mathbf{v}_0, q_0) [\varphi_1, \hat{\mathbf{v}}_2, \hat{\pi}_2] \\ &= \int_{\Omega(\phi)} \left[- \left\{ \left(\nabla \varphi_2^\top + (\nabla \varphi_2^\top)^\top - \nabla \cdot \varphi_2 \right) (\mu \nabla \mathbf{u}^\top) \right. \right. \\ & \quad + p (\nabla \cdot \varphi_2) \mathbf{I} - p (\nabla \varphi_2^\top)^\top \left. \right\} \cdot (\nabla \varphi_1^\top \nabla \mathbf{v}_0^\top) \\ & \quad - (\mu \nabla \mathbf{v}_0^\top - q_0 \mathbf{I}) \cdot \left\{ \nabla \varphi_1^\top \left(\nabla \varphi_2^\top + (\nabla \varphi_2^\top)^\top - \nabla \cdot \varphi_2 \right) \nabla \mathbf{u}^\top \right\} \\ & \quad \left. \left. + \left\{ \left((\nabla \varphi_2^\top + (\nabla \varphi_2^\top)^\top - \nabla \cdot \varphi_2) (\mu \nabla \mathbf{u}^\top) \right. \right. \right. \right. \\ & \quad \left. \left. \left. + p (\nabla \cdot \varphi_2) \mathbf{I} - p (\nabla \varphi_2^\top)^\top \right) \cdot \nabla \mathbf{v}_0^\top - q_0 \nabla \varphi_2^\top \cdot (\nabla \mathbf{u}^\top)^\top \right\} \nabla \cdot \varphi_1 \right] dx. \end{aligned} \quad (9.13.31)$$

Similarly, the third term on the right-hand side of Eq. (9.13.21) is Eq. (9.13.31) where φ_1 and φ_2 are interchanged. The fourth term on the right-hand side of Eq. (9.13.21) becomes zero.

Summarizing the above results, the second-order shape derivative of \tilde{f}_0 becomes

$$\begin{aligned} & h_0 (\phi, \mathbf{u}, \mathbf{u}) [\varphi_1, \varphi_2] \\ &= \int_{\Omega(\phi)} \left[-2 (\mu \nabla \mathbf{u}^\top \cdot \nabla \mathbf{u}^\top) (\nabla \cdot \varphi_2) (\nabla \cdot \varphi_1) \right. \\ & \quad - \left\{ (\mu \nabla \mathbf{u}^\top - p \mathbf{I}) (\nabla \mathbf{u}^\top)^\top \right\} \cdot \left\{ \nabla \varphi_2^\top \nabla \varphi_1^\top + \nabla \varphi_1^\top \nabla \varphi_2^\top \right. \\ & \quad \left. \left. + \nabla \varphi_2^\top (\nabla \varphi_1^\top)^\top + \nabla \varphi_1^\top (\nabla \varphi_2^\top)^\top \right. \right. \\ & \quad \left. \left. - 4 \nabla \varphi_2^\top \nabla \cdot \varphi_1 - 4 \nabla \varphi_1^\top \nabla \cdot \varphi_2 \right\} \right] dx. \end{aligned} \quad (9.13.32)$$

Second-Order Shape Derivative of Cost Function Using Lagrange Multiplier Method

The application of the Lagrange multiplier method in obtaining the second-order shape derivative of the mean flow resistance f_0 is described as follows. Fixing φ_1 , we define the Lagrange function for $\tilde{f}'_0(\phi)[\varphi_1] = \langle g_0, \varphi_1 \rangle$ in Eq. (9.13.18) by

$$\mathcal{L}_{10}(\phi, \mathbf{u}, p, \mathbf{w}_0, r_0) = \langle g_0, \varphi_1 \rangle - \mathcal{L}_S(\phi, \mathbf{u}, p, \mathbf{w}_0, r_0), \quad (9.13.33)$$

where \mathcal{L}_S is given by Eq. (9.13.7), and $(\mathbf{w}_0, r_0) \in U \times P$ is the adjoint variable provided for (\mathbf{u}, p) in g_0 .

Considering Eq. (9.1.6), with respect to arbitrary variations $(\varphi_2, \hat{\mathbf{u}}, \hat{p}, \hat{\mathbf{w}}_0, \hat{r}_0) \in D \times (U \times P)^2$ of $(\phi, \mathbf{u}, p, \mathbf{w}_0, r_0)$, the Fréchet derivative of \mathcal{L}_{10} is written as

$$\begin{aligned} \mathcal{L}'_{10}(\phi, \mathbf{u}, p, \mathbf{w}_0, r_0) & [\varphi_2, \hat{\mathbf{u}}, \hat{p}, \hat{\mathbf{w}}_0, \hat{r}_0] \\ &= \mathcal{L}_{10\phi'}(\phi, \mathbf{u}, p, \mathbf{w}_0, r_0) [\varphi_2] + \langle g_0(\phi), t(\varphi_1, \varphi_2) \rangle \\ &+ \mathcal{L}_{10\mathbf{u}p}(\phi, \mathbf{u}, p, \mathbf{w}_0, r_0) [\hat{\mathbf{u}}, \hat{p}] \\ &+ \mathcal{L}_{10\mathbf{w}_0r_0}(\phi, \mathbf{u}, p, \mathbf{w}_0, r_0) [\hat{\mathbf{w}}_0, \hat{r}_0]. \end{aligned} \quad (9.13.34)$$

The fourth term on the right-hand side of Eq. (9.13.34) vanishes if (\mathbf{u}, p) is the solution of the state determination problem.

The third term on the right-hand side of Eq. (9.13.34) is

$$\begin{aligned} \mathcal{L}_{10\mathbf{u}p}(\phi, \mathbf{u}, p, \mathbf{w}_0, r_0) & [\hat{\mathbf{u}}, \hat{p}] \\ &= \int_{\Omega(\phi)} \left[-2 \left\{ \left(\nabla \varphi_1^\top + (\nabla \varphi_1^\top)^\top - \nabla \cdot \varphi_1 \right) \mu \nabla \mathbf{u}^\top + p (\nabla \varphi_1^\top)^\top \right\} \cdot \nabla \hat{\mathbf{u}}^\top \right. \\ &\quad + 2 (\nabla \cdot \varphi_1) \mathbf{b} \cdot \hat{\mathbf{u}} + 2 \hat{p} (\nabla \varphi_1^\top)^\top \cdot (\nabla \mathbf{u}^\top) \\ &\quad \left. + \mu \nabla \mathbf{w}_0^\top \cdot (\nabla \hat{\mathbf{u}}^\top) - \hat{p} \nabla \cdot \mathbf{w}_0^\top - r_0 \nabla \cdot \hat{\mathbf{u}} \right] dx. \end{aligned} \quad (9.13.35)$$

Here, the conditions that Eq. (9.13.35) is zero for arbitrary $(\hat{\mathbf{u}}, \hat{p}) \in U \times P$ and \mathbf{u} satisfies the continuity equation are equivalent to setting (\mathbf{w}_0, r_0) to be the solution of the following adjoint problem.

Problem 9.13.3 (Adjoint Problem of \mathbf{w}_0 with Respect to $\langle g_0, \varphi_1 \rangle$) Under the assumption of Problem 9.13.2, let $\varphi_1 \in Y$ be given. Find $(\mathbf{w}_0, r_0) = (\mathbf{w}_0(\vartheta_1), r_0(\vartheta_1)) \in U \times P$ satisfying

$$\begin{aligned} -\nabla^\top \left(\mu \nabla \mathbf{w}_0^\top \right) + \nabla^\top r_0 &= -2 \nabla^\top \left\{ \left(\nabla \varphi_1^\top + (\nabla \varphi_1^\top)^\top - \nabla \cdot \varphi_1 \right) \mu \nabla \mathbf{u}^\top \right. \\ &\quad \left. + p (\nabla \varphi_1^\top)^\top \right\} - 2 (\nabla \cdot \varphi_1) \mathbf{b}^\top \quad \text{in } \Omega(\phi), \end{aligned}$$

$$\begin{aligned}\nabla \cdot \mathbf{w}_0 &= 2 \left(\nabla \boldsymbol{\varphi}_1^\top \right)^\top \cdot \nabla \mathbf{u}^\top \quad \text{in } \Omega(\boldsymbol{\phi}), \\ \mathbf{w}_0 &= \mathbf{0}_{\mathbb{R}^d} \quad \text{on } \partial\Omega(\boldsymbol{\phi}), \\ \int_{\Omega(\boldsymbol{\phi})} r_0 \, dx &= 0.\end{aligned}$$

□

Finally, the first and second terms on the right-hand side of Eq. (9.13.34) become

$$\begin{aligned}\mathcal{L}_{10\boldsymbol{\phi}'}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{w}_0, r_0) [\boldsymbol{\varphi}_2] &+ \langle \mathbf{g}_0(\boldsymbol{\phi}), \mathbf{t}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) \rangle \\ &= \mathcal{L}_{0\boldsymbol{\phi}'\boldsymbol{\phi}'}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{u}, p) [\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2] + \langle \mathbf{g}_0(\boldsymbol{\phi}), \mathbf{t}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) \rangle \\ &\quad - \mathcal{L}_{S\boldsymbol{\phi}'}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{w}_0, r_0) [\boldsymbol{\varphi}_2]\end{aligned}\quad (9.13.36)$$

with respect to an arbitrary $\boldsymbol{\varphi}_1 \in Y$. The first and second terms on the right-hand side of Eq. (9.13.36) are given by Eq. (9.13.25) in which $(\mathbf{v}_0, q_0) = (\mathbf{u}, p)$ is substituted. The third term is

$$\begin{aligned}-\mathcal{L}_{S\boldsymbol{\phi}'}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{w}_0, r_0) [\boldsymbol{\varphi}_2] &= \int_{\Omega(\boldsymbol{\phi})} \left[(\mu \nabla \mathbf{u}^\top - p \mathbf{I}) \cdot (\nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{w}_0^\top) + (\mu \nabla \mathbf{w}_0^\top - r_0 \mathbf{I}) \cdot (\nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{u}^\top) \right. \\ &\quad \left. + \left\{ (\mu \nabla \mathbf{u}^\top - p \mathbf{I}) \cdot \nabla \mathbf{w}_0^\top - r_0 \nabla \cdot \mathbf{u} - \mathbf{b} \cdot \mathbf{w}_0 \right\} \nabla \cdot \boldsymbol{\varphi}_2 \right] dx.\end{aligned}\quad (9.13.37)$$

Here, (\mathbf{u}, p) and (\mathbf{w}_0, r_0) are the weak solutions of Problems 9.13.1 and 9.12.3, respectively. If we denote $f_0(\boldsymbol{\phi}, \mathbf{u}, p)$ by $\tilde{f}_0(\boldsymbol{\phi})$, then we obtain the relation

$$\begin{aligned}\mathcal{L}_{10\boldsymbol{\phi}'}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{w}_0(\boldsymbol{\varphi}_1), r_0(\boldsymbol{\varphi}_1)) [\boldsymbol{\varphi}_2] &+ \langle \mathbf{g}_0(\boldsymbol{\phi}), \mathbf{t}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) \rangle \\ &= \tilde{f}_0''(\boldsymbol{\phi}) [\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2] = \langle \mathbf{g}_{H0}(\boldsymbol{\phi}, \boldsymbol{\varphi}_1), \boldsymbol{\varphi}_2 \rangle \\ &= \int_{\Omega(\boldsymbol{\phi})} \left[2\mu \left(\nabla \boldsymbol{\varphi}_1^\top \nabla \mathbf{u}^\top \right) \cdot \left(\nabla \boldsymbol{\varphi}_2^\top \nabla \mathbf{u}^\top \right) \right. \\ &\quad \left. + 2 \left\{ (\mu \nabla \mathbf{u}^\top - p \mathbf{I}) (\nabla \mathbf{u}^\top)^\top \right\} \cdot \left(\nabla \boldsymbol{\varphi}_1^\top \nabla \boldsymbol{\varphi}_2^\top \right) \right. \\ &\quad \left. - \left\{ (\mu \nabla \mathbf{u}^\top - p \mathbf{I}) (\nabla \mathbf{w}_0^\top(\boldsymbol{\varphi}_1))^\top + (\mu \nabla \mathbf{w}_0^\top(\boldsymbol{\varphi}_1) - r_0 \mathbf{I}) (\nabla \mathbf{u}^\top)^\top \right\} \right. \\ &\quad \left. \cdot \left(\nabla \boldsymbol{\varphi}_2^\top \right) \right. \\ &\quad \left. + \left\{ (\mu \nabla \mathbf{u}^\top - p \mathbf{I}) \cdot \nabla \mathbf{w}_0^\top(\boldsymbol{\varphi}_1) - \mathbf{b} \cdot \mathbf{w}_0(\boldsymbol{\varphi}_1) \right\} \nabla \cdot \boldsymbol{\varphi}_2 \right] dx,\end{aligned}\quad (9.13.38)$$

where \mathbf{g}_{H0} is the Hesse gradient of the mean flow resistance.

If $\mathbf{b} = \mathbf{0}_{\mathbb{R}^d}$ is satisfied, with respect to the solution (\mathbf{w}_0, r_0) of Problem 9.13.3,

$$\begin{aligned} \mu \nabla \mathbf{w}_0^\top (\boldsymbol{\varphi}_1) - r_0 \mathbf{I} &= 2 \left(\nabla \boldsymbol{\varphi}_1^\top + \left(\nabla \boldsymbol{\varphi}_1^\top \right)^\top - \nabla \cdot \boldsymbol{\varphi}_1 \right) \mu \nabla \mathbf{u}^\top \\ &\quad + 2p \left(\nabla \boldsymbol{\varphi}_1^\top \right)^\top, \end{aligned} \quad (9.13.39)$$

$$\nabla \mathbf{w}_0^\top (\boldsymbol{\varphi}_1) = 2 \left(\nabla \boldsymbol{\varphi}_1^\top + \left(\nabla \boldsymbol{\varphi}_1^\top \right)^\top - \nabla \cdot \boldsymbol{\varphi}_1 \right) \nabla \mathbf{u}^\top \quad (9.13.40)$$

holds. Substituting Eq. (9.13.39) and Eq. (9.13.40) into Eq. (9.13.38), and using the relation $h_0(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0)[\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2] = h_0(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0)[\boldsymbol{\varphi}_2, \boldsymbol{\varphi}_1]$, it can be confirmed that Eq. (9.13.38) accords with Eq. (9.13.32).

Shape Derivative of f_0 Using Formulae Based on Partial Shape Derivative of a Function

If the formulae based on the partial shape derivative of a function are used, the corresponding results are as follows. Here, \mathbf{b} and \mathbf{u}_D are assumed to be functions fixed in space. Moreover, it is assumed that \mathbf{u} and \mathbf{v}_0 are elements of $W^{2,2q_R}(D; \mathbb{R}^d)$, and p and q_0 are in $W^{1,2q_R}(D; \mathbb{R})$, where $q_R > d$.

Under these assumptions, the Fréchet derivative of \mathcal{L}_0 can be written as

$$\begin{aligned} \mathcal{L}'_0(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0)[\boldsymbol{\varphi}, \hat{\mathbf{u}}, \hat{p}, \hat{\mathbf{v}}_0, \hat{q}_0] &= \mathcal{L}_{0\boldsymbol{\phi}}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0)[\boldsymbol{\varphi}] \\ &\quad + \mathcal{L}_{0\mathbf{u}p}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0)[\hat{\mathbf{u}}, \hat{p}] + \mathcal{L}_{0\mathbf{v}_0q_0}(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0)[\hat{\mathbf{v}}_0, \hat{q}_0] \end{aligned} \quad (9.13.41)$$

for any arbitrary variation $(\boldsymbol{\varphi}, \hat{\mathbf{u}}, \hat{p}, \hat{\mathbf{v}}_0, \hat{q}_0) \in X \times (U \times P)^2$. Here, the notations of Eq. (9.3.21) and Eq. (9.3.27) were used. Each term is considered below.

The third term on the right-hand side of Eq. (9.13.41) accords with Eq. (9.13.12). Hence, if (\mathbf{u}, p) is the weak solution of state determination problem (Problem 9.13.1), then this term is equal to zero.

Moreover, the second term on the right-hand side of Eq. (9.13.41) is the same as Eq. (9.13.13). Hence, if the self-adjoint relationship holds, then this term is also zero.

Furthermore, the first term on the right-hand side of Eq. (9.13.41) becomes

$$\begin{aligned} \mathcal{L}_{0\boldsymbol{\phi}^*}(\boldsymbol{\phi}, \mathbf{u}, \mathbf{v}_0)[\boldsymbol{\varphi}] &= \int_{\partial\Omega(\boldsymbol{\phi})} \{ \mu \nabla \mathbf{u}^\top \cdot \nabla \mathbf{v}_0^\top - p \nabla \cdot \mathbf{v}_0 - \mathbf{b} \cdot (\mathbf{u} + \mathbf{v}_0) \} \mathbf{v} \cdot \boldsymbol{\varphi} \, d\gamma \\ &\quad + \int_{\partial\Omega(\boldsymbol{\phi})} \left[\{ (\mathbf{u} - \mathbf{u}_D) \cdot \bar{\mathbf{w}}(\boldsymbol{\varphi}, \mathbf{v}_0, q_0) + (\mathbf{v}_0 - \mathbf{u}_D) \cdot \bar{\mathbf{w}}(\boldsymbol{\varphi}, \mathbf{u}, p) \} \right. \\ &\quad \left. - \{ \mu \nabla \mathbf{u}^\top \cdot \nabla \mathbf{v}_0^\top - p \nabla \cdot \mathbf{v}_0 - \mathbf{b} \cdot (\mathbf{u} + \mathbf{v}_0) \} \mathbf{v} \cdot \boldsymbol{\varphi} \right] \, d\gamma \end{aligned}$$

$$\begin{aligned}
& + \{(\mathbf{u} - \mathbf{u}_D) \cdot (\mu \partial_v \mathbf{v}_0 - q_0 \mathbf{v}) + (\mathbf{v}_0 - \mathbf{u}_D) \cdot (\mu \partial_v \mathbf{u} - p \mathbf{v})\} (\nabla \cdot \boldsymbol{\varphi})_\tau \\
& + (\partial_v + \kappa) \{(\mathbf{u} - \mathbf{u}_D) \cdot (\mu \partial_v \mathbf{v}_0 - q_0 \mathbf{v}) \\
& + (\mathbf{v}_0 - \mathbf{u}_D) \cdot (\mu \partial_v \mathbf{u} - p \mathbf{v})\} \mathbf{v} \cdot \boldsymbol{\varphi} \} d\gamma \\
& - \int_{\partial\Omega(\boldsymbol{\phi}) \cup \Theta(\boldsymbol{\phi})} \{(\mathbf{u} - \mathbf{u}_D) \cdot (\mu \partial_v \mathbf{v}_0 - q_0 \mathbf{v}) \\
& + (\mathbf{v}_0 - \mathbf{u}_D) \cdot (\mu \partial_v \mathbf{u} - p \mathbf{v})\} \boldsymbol{\tau} \cdot \boldsymbol{\varphi} d\zeta,
\end{aligned}$$

from Eq. (9.3.21) and Eq. (9.3.27) expressing Propositions 9.3.10 and 9.3.13, where

$$\begin{aligned}
\bar{\mathbf{w}}(\boldsymbol{\varphi}, \mathbf{u}, p) = & - \left\{ (\mu \nabla \mathbf{u}^\top)^\top - p \mathbf{I} \right\} \left[\sum_{i \in \{1, \dots, d-1\}} \left\{ \boldsymbol{\tau}_i \cdot (\nabla \boldsymbol{\varphi}^\top \mathbf{v}) \right\} \boldsymbol{\tau}_i \right] \\
& + (\mathbf{v} \cdot \boldsymbol{\varphi}) \left[\nabla^\top \left\{ (\mu \nabla \mathbf{u}^\top)^\top - p \mathbf{I} \right\} \right]^\top,
\end{aligned} \quad (9.13.42)$$

and $(\nabla \cdot \boldsymbol{\varphi})_\tau$ follows Eq. (9.2.6). Here, if the Dirichlet condition of Problem 9.13.1 is considered, the terms including $\mathbf{u} - \mathbf{u}_D$ and $\mathbf{v}_0 - \mathbf{u}_D$ on $\mathcal{L}_{0\phi^*}$ vanish.

With the above results in mind, if \mathbf{u} and \mathbf{v}_0 fulfil the weak form of Problem 9.13.1 satisfying the self-adjoint relationship, we get

$$\tilde{f}'_0(\boldsymbol{\phi})[\boldsymbol{\varphi}] = \mathcal{L}_{0\phi^*}(\boldsymbol{\phi}, \mathbf{u}, \mathbf{v}_0)[\boldsymbol{\varphi}] = \langle \bar{\mathbf{g}}_0, \boldsymbol{\varphi} \rangle = \int_{\partial\Omega(\boldsymbol{\phi})} \bar{\mathbf{g}}_{\partial\Omega 0} \cdot \boldsymbol{\varphi} d\gamma \quad (9.13.43)$$

using the notation of Eq. (7.5.15) for \tilde{f}'_0 , where

$$\bar{\mathbf{g}}_{\partial\Omega 0} = \{ \mu \nabla \mathbf{u}^\top \cdot \nabla \mathbf{u}^\top - 2\mathbf{b} \cdot \mathbf{u} - 2\partial_v(\mathbf{u} - \mathbf{u}_D) \cdot (\mu \partial_v \mathbf{u} - p \mathbf{v}) \} \mathbf{v}. \quad (9.13.44)$$

Furthermore, on the homogeneous Dirichlet boundary, we have the equations

$$\nabla \mathbf{u}^\top \cdot \nabla \mathbf{u}^\top = \left\{ \mathbf{v} (\partial_v \mathbf{u})^\top \right\} \cdot \left\{ \mathbf{v} (\partial_v \mathbf{u})^\top \right\} = \partial_v \mathbf{u} \cdot \partial_v \mathbf{u} \quad (9.13.45)$$

and

$$\nabla \cdot \mathbf{u} = (\partial_v \mathbf{u}) \cdot \mathbf{v} = 0. \quad (9.13.46)$$

In this case, we get

$$\bar{\mathbf{g}}_{\partial\Omega 0} = -\mu (\partial_v \mathbf{u} \cdot \partial_v \mathbf{u}) \mathbf{v}. \quad (9.13.47)$$

From the above results, similar conclusions can be obtained for Theorem 9.8.6 with respect to function space containing $\bar{\mathbf{g}}_{\partial\Omega 0}$ of Eq. (9.13.44).

9.13.4 Relationship with Optimal Design Problem of One-Dimensional Branched Stokes Flow Field

Here, let us think about the relationship between the shape derivative of the cost function obtained with respect to the mean flow resistance minimization problem of a $d \in \{2, 3\}$ -dimensional Stokes flow field and the cross-sectional derivative of the cost function obtained with respect to the mean flow resistance minimization problem (Problem 1.3.2) of the one-dimensional branched Stokes flow field looked at in Chap. 1. To correspond to the variables and linear spaces, relationships similar to those shown in Table 9.2 with respect to the linear elastic problem hold, with an addition of the pressure to the state variable and a change of the objective function.

In Problem 1.3.2, volume force was not assumed. If this assumption is applied to Problem 9.13.2, it corresponds to setting the objective function to

$$\begin{aligned} f_0(\boldsymbol{\phi}, \mathbf{u}, p) &= \int_{\partial\Omega(\boldsymbol{\phi})} \mathbf{u}_D \cdot (\mu \partial_v \mathbf{u} - p \mathbf{v}) \, d\gamma \\ &= - \sum_{i \in \{1, 2\}} \int_0^{r_i} p_i u_{Hi}(r) 2\pi r \, dr, \end{aligned} \quad (9.13.48)$$

where p_i and r_i for $i \in \{0, 1, 2\}$ follows the definition given in Problem 1.3.2, respectively, and u_{Hi} is given by Eq. (1.3.1). In Eq. (9.13.48), $\partial_v \mathbf{u} = 0$ and $p_0 = 0$ were used. In Problem 1.3.2, u_1 and u_2 were defined as the volumes of the fluid flow per unit time on Γ_1 and Γ_2 , respectively, and were fixed during the changes on the cross-sectional areas. In other words, u_{H1} and u_{H2} were varying with the boundary measure. This relation is written as

$$u'_{Hi}(r)[\boldsymbol{\varphi}] = -u_{Hi}(r)(\nabla \cdot \boldsymbol{\varphi}_i)_\tau. \quad (9.13.49)$$

On the other hand, \mathbf{u}_D was taken to be fixed with material in Eq. (9.13.18) which gives the shape derivative of f_0 . If this difference is considered, the shape derivative of f_0 defined by Eq. (9.13.48) becomes

$$\begin{aligned} \tilde{f}'_0(\boldsymbol{\phi})[\boldsymbol{\varphi}] &= \langle \mathbf{g}_0, \boldsymbol{\varphi} \rangle \\ &= \sum_{i \in \{0, 1, 2\}} l \int_{\Gamma_i} \left(\mathbf{G}_{\Omega 0i} \cdot \nabla \boldsymbol{\varphi}_i^\top + g_{\Omega 0i} \nabla \cdot \boldsymbol{\varphi}_i \right) \, d\gamma \\ &\quad + \sum_{i \in \{1, 2\}} \int_0^{r_i} p_i u_{Hi}(r) (\nabla \cdot \boldsymbol{\varphi}_i)_\tau 2\pi r \, dr. \end{aligned} \quad (9.13.50)$$

We express a point in the cylindrical domain in the one-dimensional Stokes flow field as $(x, r, \theta) \in (0, l) \times \Gamma_i$ and $\mathbf{u} = (u_{Hi}(r), 0, 0)^\top$. Here, because of the relationship

$$\int_{\Gamma_i} (\nabla \cdot \boldsymbol{\varphi}_i)_\tau \, d\gamma = (\nabla \cdot \boldsymbol{\varphi}_i)_\tau a_i = b_i,$$

the following equations hold:

$$\nabla \cdot \boldsymbol{\varphi}_i = (\nabla \cdot \boldsymbol{\varphi}_i)_\tau = \frac{b_i}{a_i}, \quad \nabla \boldsymbol{\varphi}_i^\top = \begin{pmatrix} 0 & 0 & 0 \\ 0 & b_i/a_i & 0 \\ 0 & 0 & 0 \end{pmatrix}. \quad (9.13.51)$$

The flow velocity is given by

$$\nabla \mathbf{u}^\top = \frac{p_i - \bar{p}}{4\mu l} \begin{pmatrix} 0 & 0 & 0 \\ 2r & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Hence, we have

$$\begin{aligned} \mathbf{G}_{\Omega 0i} \cdot \nabla \boldsymbol{\varphi}_i^\top &= -2 \left\{ \left(\mu \nabla \mathbf{u}^\top - p \mathbf{I} \right) \left(\nabla \mathbf{u}^\top \right)^\top \right\} \cdot \left(\nabla \boldsymbol{\varphi}_i^\top \right) \\ &= -8\mu \left(\frac{p_i - \bar{p}}{4\mu l} \right)^2 \frac{b_i}{a_i} r^2, \\ g_{\Omega 0i} \nabla \cdot \boldsymbol{\varphi}_i &= \mu \nabla \mathbf{u}^\top \cdot \nabla \mathbf{u}^\top (\nabla \cdot \boldsymbol{\varphi}_i) = 4\mu \left(\frac{p_i - \bar{p}}{4\mu l} \right)^2 \frac{b_i}{a_i} r^2. \end{aligned}$$

From these results and Eq. (1.3.2), we get

$$\begin{aligned} \tilde{f}'_0(\boldsymbol{\phi})[\boldsymbol{\varphi}] &= \sum_{i \in \{0, 1, 2\}} l \int_0^{r_i} \left(\mathbf{G}_{\Omega 0i} \cdot \nabla \boldsymbol{\varphi}_i^\top + g_{\Omega 0i} \nabla \cdot \boldsymbol{\varphi}_i \right) 2\pi r \, dr \\ &\quad - \sum_{i \in \{1, 2\}} p_i \frac{u_i b_i}{a_i} \\ &= - \sum_{i \in \{0, 1, 2\}} 2 \frac{u_i^2 b_i}{a_i^3} = \mathbf{g}_0 \cdot \mathbf{b}. \end{aligned} \quad (9.13.52)$$

In Eq. (9.13.52), the equation of continuity with respect to domain variation

$$\sum_{i \in \{0, 1, 2\}} \int_{\Gamma_i} u_i (\nabla \cdot \boldsymbol{\varphi}_i)_\tau \, d\gamma = \frac{u_0 b_0}{a_0} + \frac{u_1 b_1}{a_1} + \frac{u_2 b_2}{a_2} = 0 \quad (9.13.53)$$

was used. Here, \mathbf{g}_0 matches the cross-sectional-area gradient of the mean flow resistance f_0 with respect to the one-dimensional branched Stokes flow field obtained in Eq. (1.3.19).

Furthermore, the Hessian form of f_0 becomes

$$\begin{aligned}
 h_0(\boldsymbol{\phi}, \mathbf{u}, p, \mathbf{v}_0, q_0) & [\boldsymbol{\varphi}_{1i}, \boldsymbol{\varphi}_2] \\
 &= \sum_{i \in \{1, 2\}} l \int_{\Gamma_i} \left[-2 \left(\mu \nabla \mathbf{u}^\top \cdot \nabla \mathbf{u}^\top \right) (\nabla \cdot \boldsymbol{\varphi}_{2i}) (\nabla \cdot \boldsymbol{\varphi}_{1i}) \right. \\
 &\quad \left. - \left\{ \left(\mu \nabla \mathbf{u}^\top - p \mathbf{I} \right) (\nabla \mathbf{u}^\top)^\top \right\} \cdot \left\{ \nabla \boldsymbol{\varphi}_{2i}^\top \nabla \boldsymbol{\varphi}_{1i}^\top + \nabla \boldsymbol{\varphi}_{1i}^\top \nabla \boldsymbol{\varphi}_{2i}^\top \right. \right. \\
 &\quad \left. \left. + \nabla \boldsymbol{\varphi}_{2i}^\top (\nabla \boldsymbol{\varphi}_{1i}^\top)^\top + \nabla \boldsymbol{\varphi}_{1i}^\top (\nabla \boldsymbol{\varphi}_{2i}^\top)^\top \right. \right. \\
 &\quad \left. \left. - 4 \nabla \boldsymbol{\varphi}_{2i}^\top \nabla \cdot \boldsymbol{\varphi}_{1i} - 4 \nabla \boldsymbol{\varphi}_{1i}^\top \nabla \cdot \boldsymbol{\varphi}_{2i} \right\} \right] d\gamma \\
 &\quad + \sum_{i \in \{0, 1, 2\}} \left\{ \frac{d}{da_i} \left(\frac{u_i^2 b_{1i}}{a_i^3} \right) b_{2i} + \frac{u_i^2 b_{1i}}{a_i^3} \left(\frac{b_{2i}}{a_i} \right) \right\} \\
 &= \sum_{i \in \{0, 1, 2\}} 6 \frac{u_i^2}{a_i^4} b_{1i} b_{2i} = \mathbf{b}_1 \cdot (\mathbf{H}_0 \mathbf{b}_2) \tag{9.13.54}
 \end{aligned}$$

using Eq. (9.13.32). Here, \mathbf{H}_0 matches the Hessian matrix of the mean flow resistance f_0 with respect to the cross-sectional areas of the one-dimensional branched Stokes flow field obtained in Eq. (1.3.26).

9.13.5 Numerical Example

The results of mean flow resistance minimization for a two-dimensional Stokes flow field around an isolated object are shown in Figs. 9.24, 9.25, 9.26, 9.27. The boundary condition of the state determination problem is assumed to be a uniform flow field in the horizontal direction on the outer boundary and zero on the boundary of the isolated object as shown in Fig. 9.24a. Moreover, with respect to the boundary condition for domain variation, the outer boundary was fixed (added in $\bar{\Omega}_{C0}$ of Eq. (9.1.1)). The programs were written using the programming language FreeFEM (<https://freefem.org/>) [66] for the finite element method. In the finite element analyses of the Stokes problem, triangular elements of the second order with respect to the velocity and of the first order with respect to the pressure were used. Also, in the finite element analyses of the H^1 gradient method or the H^1 Newton method, the second-order triangular elements were used. On the other hand, in the case using the H^1 Newton method, the routine for the second-order method

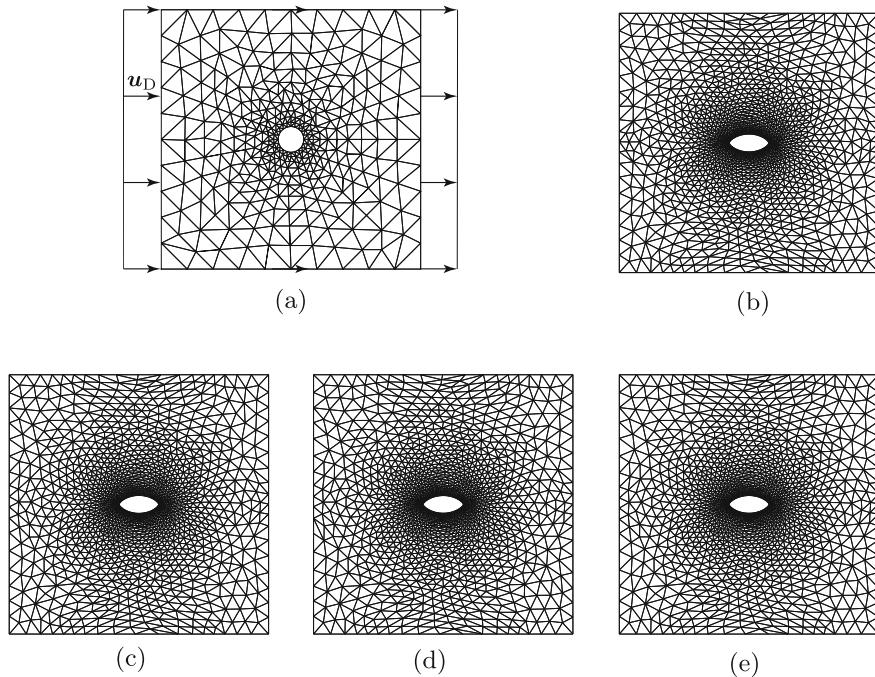


Fig. 9.24 Numerical example of mean flow resistance minimization problem: shape ($k = 40$). (a) Initial shape and boundary condition. (b) H^1 gradient method ($\bar{g}_{\mathcal{L}}$). (c) H^1 gradient method ($g_{\mathcal{L}}$). (d) H^1 Newton method ($h_0, g_{\mathcal{L}}$). (e) H^1 Newton method ($g_{H0}, g_{\mathcal{L}}$)

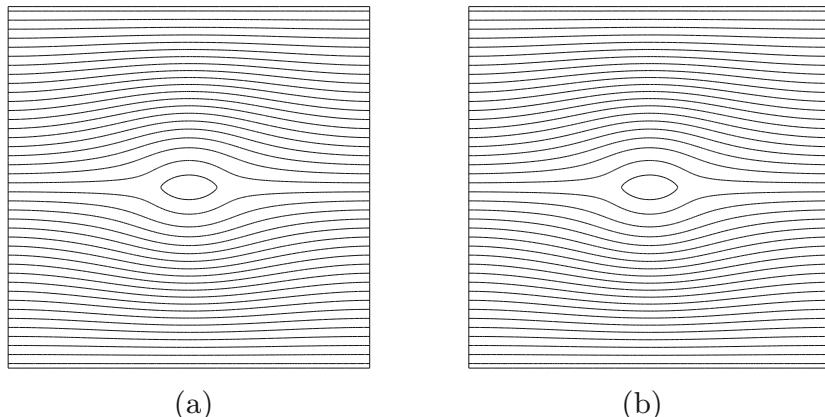


Fig. 9.25 Numerical example of mean flow resistance minimization problem: streamlines. (a) Streamline of initial shape. (b) Streamline of optimal shape (H^1 Newton method, $k = 40$)

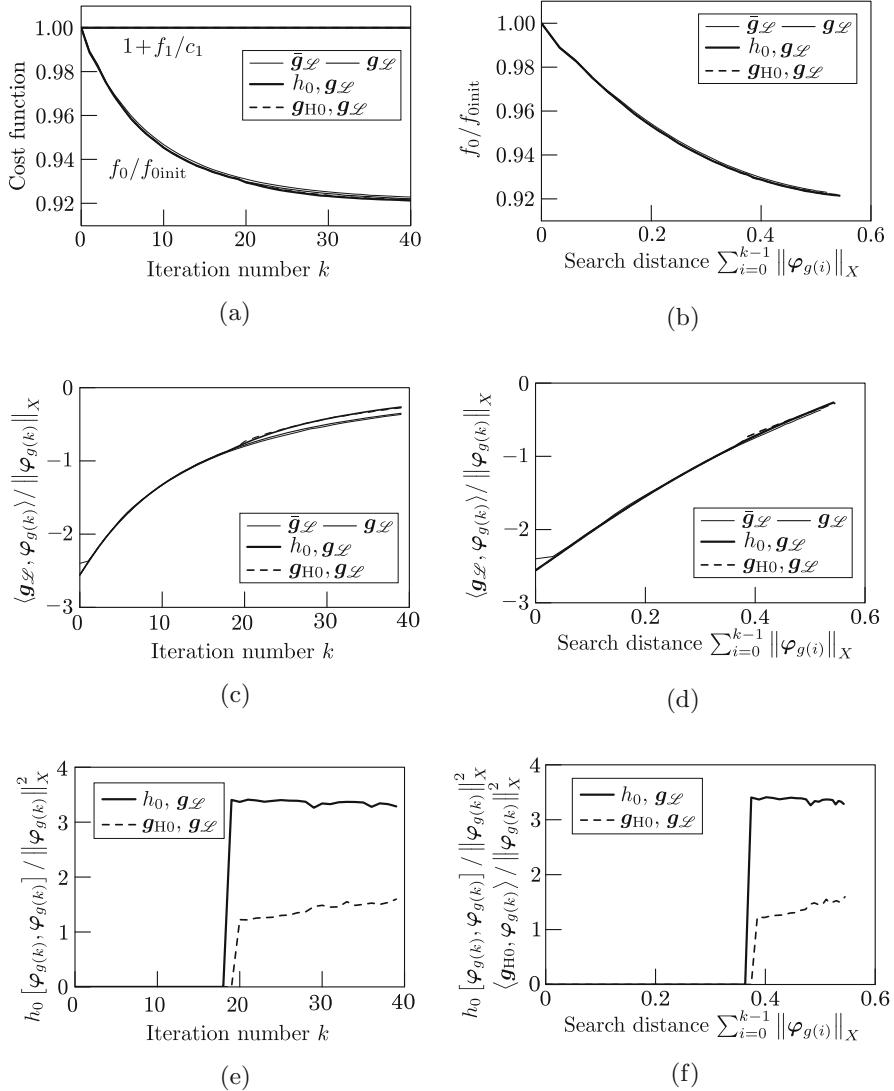


Fig. 9.26 Numerical example of mean flow resistance minimization problem: cost functions, their gradients and Hessians on the search path ($\bar{g}_{\mathcal{L}}$: H^1 gradient method using $\bar{g}_{\mathcal{L}}$, $g_{\mathcal{L}}$: H^1 gradient method using $g_{\mathcal{L}}$, $h_0, g_{\mathcal{L}}$: H^1 Newton method, $g_{H0}, g_{\mathcal{L}}$: H^1 Newton method using Hesse gradient). (a) Cost functions. (b) Cost functions (search distance). (c) Gradient of f_0 on search path. (d) Gradient of f_0 on search path (search distance). (e) Hessian of f_0 on search path. (f) Hessian of f_0 on search path (search distance)

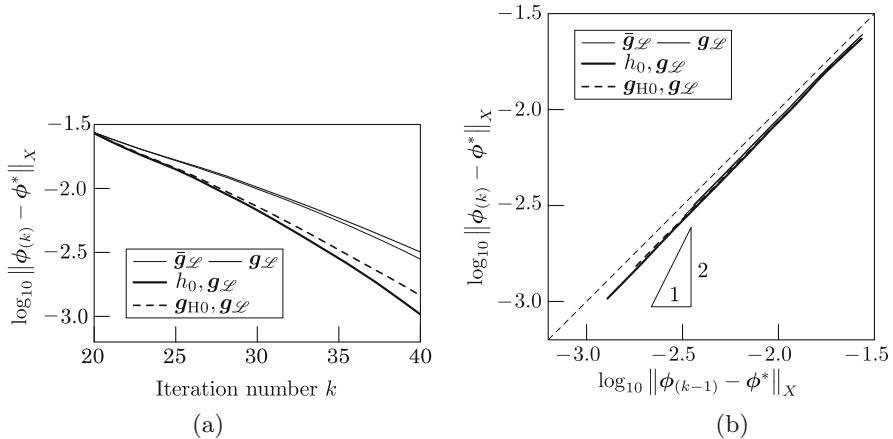


Fig. 9.27 Numerical example of mean flow resistance minimization problem: distance $\|\phi_{(k)} - \phi^*\|_X$ from an approximate minimum point ϕ^* (\bar{g}_L : the gradient method, h_L, g_L : the Newton method, g_{H0}, h_1, g_L : the Newton method using the Hesse gradient). (a) Iteration history. (b) $(k-1)$ -th vs. k -th plot

was started at $k_N = 20$. The parameters (c_a in Eq. (9.10.1), c_Ω in Eq. (9.9.3), k_N , $c_{\Omega 1}$ and $c_{\Omega 0}$ in Eq. (9.9.17), c_h in Eq. (9.10.8) and the parameter (*errelas*) that controls the error level in the adaptive mesh) affect the result. For a complete understanding of the conditions, we suggest that the readers also examine the details of the programs.⁵

Figure 9.24b–e show the shapes obtained by the four methods (H^1 gradient method using $\bar{g}_L = \bar{g}_0 + \lambda_1 \bar{g}_1$ of the boundary integral type, H^1 gradient method using $g_L = g_0 + \lambda_1 g_1$ of the domain integral type, H^1 Newton method using h_0 and g_L , and H^1 Newton method using g_{H0} and g_L). Figures 9.25a,b illustrate the streamlines in the initial shape and the optimal shape obtained by the H^1 Newton method, respectively. The streamlines are defined as the contour lines of the flow function $\psi : \Omega(\phi) \rightarrow \mathbb{R}$ when the flow velocity \mathbf{u} is given by $(\partial\psi/\partial x_2, -\partial\psi/\partial x_1)^\top$.

The graphs in Fig. 9.26 illustrate the histories of the cost functions and the gradients and Hessians of the object function f_0 on the search path with respect to the iteration number k and the search distance $\sum_{i=0}^{k-1} \|\varphi_{g(i)}\|_X$. In this figure, f_0 denotes the value of f_0 at the initial density. Also, c_1 is set as the integral volume. The gradient of f_0 on the search path was calculated using the Lagrange function $\mathcal{L} = \mathcal{L}_0 + \lambda_1 f_1$ by $\langle g_L, \varphi_{g(k)} \rangle / \|\varphi_{g(k)}\|_X$. The Hessian of f_0 on the search path was computed by $h_0 [\varphi_{g(k)}, \varphi_{g(k)}] / \|\varphi_{g(k)}\|_X^2$. In the case of the Newton method using the Hesse gradient, the formula $\langle g_{H0}, \varphi_{g(k)} \rangle / \|\varphi_{g(k)}\|_X^2$ was used to calculate

⁵See Electronic supplementary material.

the Hessian. The norm $\|\boldsymbol{\varphi}_{g(i)}\|_X$ of the i -th search vector is defined by Eq. (9.12.67). The computational times until $k = 40$ by PC were 16.324, 43.628, 63.173, 81.039 s by the H^1 gradient method of the boundary integral type, the H^1 gradient method of the domain integral type, the H^1 Newton method and the H^1 Newton method using the Hesse gradient, respectively.

Regarding the computational results obtained from the above-mentioned numerical illustrations, our findings were similar to those given in Sect. 9.12.5. The graphs in Fig. 9.26a show that the convergence speed with respect to the iteration number k is faster when using the H^1 Newton method than when employing the H^1 gradient method. However, we emphasize that $c_{\Omega 1}$ and $c_{\Omega 0}$ in Eq. (9.9.17) were actually replaced with smaller values (the step size was enlarged) within the area where numerical instability did not happen when the H^1 Newton method started, and it seems that the convergence speed was improved due to the increased in the step size. Moreover, the aspect around the minimum point can be observed in Fig. 9.26d,f. From these graphs, based on fact that the Hessian of f_0 on the search path is positive valued, we infer that the point of convergence is a local minimum point.

In addition, Fig. 9.27a shows the graphs of the distance $\|\boldsymbol{\phi}_{(k)} - \boldsymbol{\phi}^*\|_X$ of the k -th approximate $\boldsymbol{\phi}_{(k)}$ obtained by the four methods to the approximate minimum point $\boldsymbol{\phi}^*$ with respect to the iteration number k . The approximate minimum point $\boldsymbol{\phi}^*$ is substituted by the numerical solution of $\boldsymbol{\phi}$ when the iteration time is taken larger than the given value in the H^1 Newton method. From this figure, it can easily be observed that the convergence orders for the results obtained via the H^1 Newton methods are higher than the first order. However, from Fig. 9.27b, which shows the plot of the k -th distance $\|\boldsymbol{\phi}_{(k)} - \boldsymbol{\phi}^*\|_X$ with respect to the $(k-1)$ -th distance, it can be observed easily that the convergence order of the H^1 Newton method is less than the second order but is definitely more than the first order. The reason behind this result is considered as the same as that stated at the end of Sect. 9.12.5.

9.14 Summary

In Chap. 9, a shape optimization problem of domain variation type was constructed with respect to the domain on which a boundary value problem of a partial differential equation is defined and its solution looked at in detail. The key points are as below:

- (1) In a shape optimization problem of domain variation type, the design variable is a function defined on an initial domain and represents the displacement of each of the points from the reference domain to the new domain after variation (Sect. 9.1). The linear space X and the admissible set \mathcal{D} of the design variable are defined by Eqs. (9.1.1) and (9.1.3), respectively. Furthermore, in Sect. 9.1.3, two notions of derivatives called the shape derivative and partial shape derivative were introduced with respect to a function and a functional defined on varying domains.

- (2) In Sect. 9.2, the formulae for the shape derivatives relating to the Jacobi matrix of domain mapping were established. Using these formulae, it was shown in Sect. 9.3 that the shape derivatives of functions and functionals can be obtained and there corresponding forms were established. These formulae, in addition, were used to define a variety of variation rules for functions in Sect. 9.4.
- (3) In Sect. 9.6, considering a Poisson problem as a state determination problem (Sect. 9.5), a shape optimization problem of domain variation type was defined on X .
- (4) It was shown that the shape derivative of a cost function can be obtained via the Lagrange multiplier method. In this case, an evaluation method using the formulae based on the shape derivative of a function given in Theorem 9.8.2 and evaluation method using the formulae based on the shape derivative of a function stated in Theorem 9.8.6 can be considered. However, these shape derivatives are not necessarily in the linear space containing the admissible set for the design variables, as pointed out in Remark 9.8.7.
- (5) In Sect. 9.9, an H^1 gradient method using the shape derivative of a cost function was defined on the space X . The solutions of the H^1 gradient method are contained in the admissible set (Theorem 9.9.6) excluding the neighborhoods of singular points. Furthermore, in Sect. 9.9.2, it was shown that if the second-order shape derivative of the cost function can be calculated, it is also possible to obtain a descent direction for the cost function using the H^1 Newton method.
- (6) In Sect. 9.10, it was shown that a solution to a shape optimization problem of domain variation type with constraints can be constructed using the same framework as the gradient method with respect to constrained problems and Newton method with respect to constrained problems shown in Chap. 3.
- (7) When the finite element method is used to solve a state determination problem, the adjoint problem with respect to f_i and the H^1 gradient method, the order evaluation of the finite element solution with respect to the search vector φ_g can be obtained (Theorem 9.11.5).
- (8) In Sect. 9.12, the first and second order shape derivatives of some cost functions associated with a mean compliance minimization problem of a linear elastic body with domain measure constraint were established.
- (9) In Sect. 9.13, the first and second-order shape derivatives of some cost functions associated with a mean flow resistance minimization problem of a Stokes flow field with domain measure constraint were established.

Formulations and solutions of certain topology optimization problems of density variation type and shape optimization problems of domain variation type were introduced in Chaps. 8 and 9, respectively. As concluding remarks, we give a comparison of these problems below, detailing their advantages and disadvantages.

In the case of the density variation type, the density defined on a fixed domain bears advantages and disadvantages. An advantage is that clear theoretical development could easily be carried out because it enters the conventional framework of a typical function optimization problem. Moreover, replacing the density of the design variable by other material parameters, various problems except the topology

optimization problem can be formulated. For example, when we use a healthy rate of stiffness instead of the density, an identification problem of damage in a linear elastic body can be constructed [160]. On the other hand, a disadvantage can be mentioned due to the need of some additional scheme to determine the boundary of a continuum from the obtained density.

In contrast, in the case of the domain variation type, it is necessary to prepare various formulas to obtain the shape derivatives of cost functions. This is primarily due to the fact that the domains where the associated state determination problems were defined vary. Especially, when one calculates the second-order derivative, it is not easy to notice that a correction term (refer to Eqs. (9.1.9) and (9.3.11)) proportional to the first-order shape derivative of the cost function that was obtained with respect to the product of the second variation vector and variation of the first perturbation vector. On the other hand, it is possible to perturb a boundary of a continuum directly in actual numerical analysis. Hence, it is superior in the sense that the shape can be found correctly.

In actual shape optimization problems, it is hoped that a suitable method could easily be chosen considering its desired features.

9.15 Practice Problems

9.1 Suppose condition (2) in Theorem 9.8.2 holds. Show that the second term on the right-hand side of Eq. (9.8.9) giving \mathbf{g}_{pi} is in $L^\infty(\Gamma_p(\phi); \mathbb{R}^d)$.

9.2 In Problems 9.5.4 and 9.12.1, which were considered as state determination problems in this chapter, a mixed Dirichlet–Neumann boundary condition was assumed. However, in order to obtain the results in Theorem 9.8.2, Hypothesis 9.5.3 (2) ($\beta < \pi/3$ when on mixed boundaries) has to be satisfied with respect to the opening angle β . If the mixed boundary condition is replaced with a Robin condition, Hypothesis 9.5.3 (1) ($\beta < 2\pi/3$ when on boundaries of the same type) then becomes applicable. Hence, if the extended Poisson problem taken up in Chap. 5 (Problem 5.1.3) is simplified by removing the terms unrelated to the boundary conditions and replacing with a domain variation type, then we obtain the following.

Problem 9.15.1 (Poisson Problem of Robin Type) Let $\phi \in \mathcal{D}$ and $c_{\partial\Omega}(\phi) : D \rightarrow \mathbb{R}$ and $p_R(\phi) : D \rightarrow \mathbb{R}$ be given functions fixed with the material. Find $u : \Omega(\phi) \rightarrow \mathbb{R}$ which satisfies

$$\begin{aligned} -\Delta u &= 0 && \text{in } \Omega(\phi), \\ \partial_\nu u + c_{\partial\Omega}(\phi)u &= p_R(\phi) && \text{on } \partial\Omega(\phi). \end{aligned} \quad \square$$

Here, choose Problem 9.15.1 as the state determination problem and let the cost function be

$$f_i(\phi, u) = \int_{\partial\Omega(\phi)} \eta_{Ri}(\phi, u) d\gamma \quad (9.15.1)$$

for $i \in \{0, 1, \dots, m\}$, where $\eta_{Ri}(\phi, u)$ is some function fixed with the material. In this case, compute the shape derivative g_i using the formulae based on the shape derivative of a function. Moreover, state the condition of the corner opening angle and the required regularities for $c_{\partial\Omega}$, p_R , η_{Ri} and η_{Riu} in order to have a similar regularity result for g_i in Theorem 9.8.2.

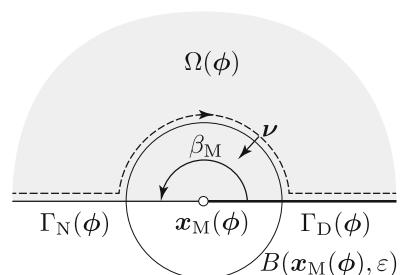
9.3 In a shape optimization problem of domain variation type examined in this chapter, if there is a crack on $\partial\Omega(\phi)$ (opening angle $\beta = 2\pi$), or if there is a Dirichlet boundary and Neumann boundary on a smooth boundary (opening angle $\beta = \pi$), then Hypothesis 9.5.3 is not satisfied. This would then imply that the assumption $u \in \mathcal{S}$ in Theorem 9.8.2 is also not satisfied and therefore, it is not clear that the shape derivative is obtained as an element of X' . However, if the linear space of the design variable (domain variation) is replaced with

$$X = \left\{ \phi \in C^{0,1}(D; \mathbb{R}^d) \mid \phi = \mathbf{0}_{\mathbb{R}^d} \text{ on } \bar{\Omega}_{C0} \right\},$$

it is possible to show that the shape derivative of the corresponding cost function is a bounded linear functional with respect to this space. The shape derivative can then be computed using a generalized J integral [12].

Here, let us go through the calculation of an interim result used in obtaining the shape derivative. Suppose $\Omega(\phi)$ is a two-dimensional domain and x_C is the tip of a crack (opening angle $\beta_C = 2\pi$) and of an interior point on the homogeneous Dirichlet boundary or homogeneous Neumann boundary. Moreover, suppose x_M , as shown in Fig. 9.28, is a point on a smooth boundary and a boundary between the Dirichlet and Neumann boundaries (opening angle $\beta_M = \pi$). In this case, think about obtaining the shape derivative of a cost function at x_C and x_M with the corresponding state determination problem defined as follows.

Fig. 9.28 Path of boundary integral \mathcal{P}_u



Problem 9.15.2 (Poisson Problem of Domain Variation Type) Let $\phi \in \mathcal{D}$ and $b(\phi)$ be a given function fixed with the material. Find $u : \Omega(\phi) \rightarrow \mathbb{R}$ which satisfies

$$\begin{aligned} -\Delta u &= b(\phi) \quad \text{in } \Omega(\phi), \\ \partial_\nu u &= 0 \quad \text{on } \Gamma_N(\phi), \\ u &= 0 \quad \text{on } \Gamma_D(\phi) \end{aligned}$$

□

Here, we replace the cost function of Eq. (9.6.1) with

$$f_i(\phi, u) = \int_{\Omega(\phi)} \zeta_i(\phi, u) \, dx - c_i,$$

and assume, for simplicity, that ζ_i is not a function of ∇u . The shape derivative of f_i is computed as

$$\langle g_i, \varphi \rangle = -\mathcal{P}_u(\partial\Omega(\phi), \varphi, u) [v_i] + \langle \hat{g}_{iC}, \varphi \rangle + \langle \hat{g}_{iM}, \varphi \rangle + \langle g_{iR}, \varphi \rangle$$

using the \mathcal{P} integral defined in a generalized J integral [12], where

$$\begin{aligned} -\mathcal{P}_u(\partial\Omega(\phi), \varphi, u) [v_i] &= \int_{\partial\Omega(\phi)} \{ (\nabla u \cdot \nabla v_i) \, \mathbf{v} \cdot \varphi \\ &\quad - \partial_\nu u \nabla v_i \cdot \varphi - \partial_\nu v_i \nabla u \cdot \varphi \} \, d\gamma, \end{aligned} \quad (9.15.2)$$

$$\begin{aligned} \langle \hat{g}_{ij}, \varphi \rangle &= \lim_{\epsilon \rightarrow 0} - \int_0^{\beta_j} \{ (\nabla u \cdot \nabla v_i) \, \mathbf{v} \cdot \varphi - \partial_\nu u \nabla v_i \cdot \varphi \\ &\quad - \partial_\nu v_i \nabla u \cdot \varphi \} \epsilon \, d\theta, \end{aligned} \quad (9.15.3)$$

$$\langle g_{iR}, \varphi \rangle = \int_{\partial\Omega(\phi)} b v_i \, \mathbf{v} \cdot \varphi \, d\gamma + \int_{\Omega(\phi)} (\zeta_i \phi \cdot \varphi + \zeta_i \nabla \cdot \varphi) \, dx, \quad (9.15.4)$$

for $j \in \{C, M\}$, and $\beta_C = 2\pi$ and $\beta_M = \pi$ with respect to. Here, Eq. (9.15.3) is the shape derivative of f_i with respect to the variation of the singular point. u and v_i are given by

$$u(r, \theta) = k_j r^{\pi/\beta_j} \cos \frac{\pi}{\beta_j} \theta + u_R, \quad (9.15.5)$$

$$v_i(r, \theta) = l_{ij} r^{\pi/\beta_j} \cos \frac{\pi}{\beta_j} \theta + v_{iR} \quad (9.15.6)$$

using (r, θ) coordinate with x_j as the origin with respect to $j \in \{C, M\}$ as seen in Sect. 5.3. Here, k_j and l_{ij} are constants and u_R and v_{iR} are elements of $H^2(D; \mathbb{R})$. In this case, substitute Eqs. (9.15.5) and (9.15.6) into Eq. (9.15.3) and obtain \hat{g}_{iC} and \hat{g}_{iM} .

Appendices

A.1 Basic Terminology

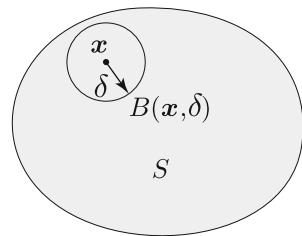
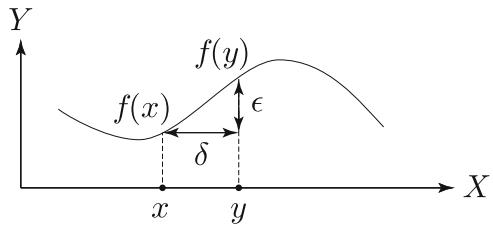
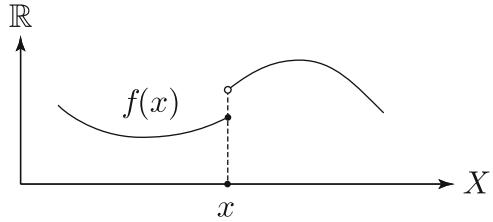
Firstly, let us define the basic terminology used in mathematics.

A.1.1 *Open Sets, Closed Sets and Bounded Sets*

Let X be a metric space (Definition 4.2.8), for example \mathbb{R}^2 , and S be its subset. We define the distance between two elements in X by $d : X \times X \rightarrow \mathbb{R}$. For $x \in X$, the open ball $B(x, \delta) = \{y \in X \mid d(x, y) < \delta\}$ of radius $\delta > 0$ is called a neighborhood. A subset S is called an open set if, for an arbitrary $x \in S$, there exists $\delta > 0$ such that $B(x, \delta) \subset S$ (Fig. A.1). If $X \setminus S$ is an open set, S is said to be a closed set. The smallest closed set containing S is its closure and is denoted by \bar{S} . For an open set S , $\partial S = \bar{S} \setminus S$ is referred to as the boundary of S . $x \in S$ is called an inner point and $x \in \partial S$ a boundary point. For a closed set S , the set of all inner points is called the interior and is denoted by S° . Moreover, a subset S is a bounded set if there exists an arbitrary point x of X and $\beta > 0$ such that for every $y \in S$, $d(x, y) \leq \beta$ holds.

A.1.2 *Continuity of Functions*

The continuity of a function on a metric space is defined as follows. Let X and Y be metric spaces with corresponding distances $d_X : X \times X \rightarrow \mathbb{R}$ and $d_Y : Y \times Y \rightarrow \mathbb{R}$. A function $f : X \rightarrow Y$ is said to be continuous on $x \in X$ if, for every $\epsilon > 0$, there

Fig. A.1 Open set S **Fig. A.2** Continuity of a function**Fig. A.3** Lower semi-continuity of a function

exists $\delta > 0$ such that

$$d_Y(f(y), f(x)) < \epsilon$$

holds for any $y \in X$ that satisfies $d_X(x, y) < \delta$ (Fig. A.2). Moreover, if it is continuous on all $x \in X$, it is called uniformly continuous.

From this definition, a uniformly continuous function is continuous but its converse is not necessarily true. Take, for instance, $f(x) = x^2$. This function is clearly continuous on \mathbb{R} . However, as $|x| \rightarrow \infty$, the gradient quickly becomes large. In this case, $\delta \rightarrow 0$ for any given $\epsilon > 0$, contradicting the requirement that we can find a $\delta > 0$. Nevertheless, a continuous function is also uniformly continuous and bounded if its domain is a compact (bounded and closed) set.

Moreover, an extended real-valued function $f : X \rightarrow \mathbb{R} \cup \{-\infty, \infty\}$ is said to be lower semi-continuous on $x \in X$ if, for every $\epsilon > 0$, there exists $\delta > 0$ such that

$$f(y) \geq f(x) - \epsilon$$

holds for any $y \in X$ that satisfies $d_X(x, y) < \delta$ (Fig. A.3).

A.2 Positive Definiteness of Real Symmetric Matrix

In Chaps. 2 and 3, the positive definite real symmetric matrix plays an important role. Here, let us state a fundamental theorem used to check the positive definiteness of a matrix based on the notion of eigenvalues.

Theorem A.2.1 (Positive Definiteness of a Real Symmetric Matrix) *Let $A \in \mathbb{R}^{d \times d}$ be a real symmetric matrix. The necessary and sufficient conditions for A to be positive definite is that all of its eigenvalues $\lambda_1, \dots, \lambda_d \in \mathbb{R}$ are positive.* \square

Proof Let us show the necessity part. For each $i \in \{1, \dots, d\}$, let $\mathbf{x}_i \in \mathbb{R}^d$ be the eigenvector corresponding to the eigenvalue λ_i . If A is positive definite, the condition

$$\mathbf{x}_i \cdot A\mathbf{x}_i = \mathbf{x}_i \cdot (\lambda_i \mathbf{x}_i) = \lambda_i \|\mathbf{x}_i\|^2 > 0$$

holds for every $i \in \{1, \dots, d\}$. It follows that $\lambda_i > 0$ for all $i \in \{1, \dots, d\}$. For the sufficiency part, it can be shown that \mathbf{x}_i , for $i \in \{1, \dots, d\}$, are mutually orthogonal. An arbitrary vector $\mathbf{x} \in \mathbb{R}^d$ is given by a linear combination of d independent vectors. That is, we can write

$$\mathbf{x} = \sum_{i \in \{1, \dots, d\}} \mathbf{x}_i \xi_i$$

and from this, we have

$$\mathbf{x} \cdot A\mathbf{x} = \sum_{i \in \{1, \dots, d\}} \lambda_i \|\mathbf{x}_i\|^2 \xi_i^2 > 0$$

for all $\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}_{\mathbb{R}^d}\}$. \square

A basis for evaluating positive definiteness based on minors is known as Sylvester's criterion (see, e.g., [54]).

Theorem A.2.2 (Sylvester's Criterion) *Let $A = (A_{ij})_{ij} \in \mathbb{R}^{d \times d}$ be a real symmetric matrix. The necessary and sufficient condition for A to be positive definite is that all minors with respect to $i \in \{1, \dots, d\}$ satisfy*

$$|A_i| = \begin{vmatrix} A_{11} & \cdots & A_{1i} \\ \vdots & \ddots & \vdots \\ A_{i1} & \cdots & A_{ii} \end{vmatrix} > 0. \quad \square$$

A.3 Null Space, Image space and Farkas's Lemma

In Chap. 2, the relationship between null space and image space as well as Farkas's lemma are used in the proofs of some important theorems. Here, let us summarize these auxiliary results.

Let m and n be natural numbers and $\mathbf{A} = (\mathbf{a}_1 \ \mathbf{a}_2 \ \cdots \ \mathbf{a}_m) \in \mathbb{R}^{n \times m}$. The set

$$\text{Ker } \mathbf{A} = \{ \mathbf{y} \in \mathbb{R}^m \mid \mathbf{A}\mathbf{y} = \mathbf{0}_{\mathbb{R}^n} \} \quad (\text{A.3.1})$$

is called the null space or the kernel space of \mathbf{A} . Moreover,

$$\text{Im } \mathbf{A} = \{ \mathbf{A}\mathbf{z} \in \mathbb{R}^n \mid \mathbf{z} \in \mathbb{R}^m \} \quad (\text{A.3.2})$$

is called the image space or the range space of \mathbf{A} . On the other hand, when a linear subspace V of \mathbb{R}^n is given, the linear space constructed of all vectors orthogonal to V is referred to as the orthogonal complement space of V and is written as V^\perp . The spaces $\text{Ker } \mathbf{A}$ and $\text{Im } \mathbf{A}$ are related as follows.

Lemma A.3.1 (Orthogonal Complement of Null Space and Image Space) *For a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$, the following relations hold:*

$$\text{Im } \mathbf{A} = (\text{Ker } \mathbf{A}^\top)^\perp, \quad (\text{Im } \mathbf{A}^\top)^\perp = \text{Ker } \mathbf{A}. \quad \square$$

Proof By definition, we have

$$\begin{aligned} \text{Ker } \mathbf{A}^\top &= \left\{ \mathbf{y} \in \mathbb{R}^n \mid \mathbf{A}^\top \mathbf{y} = \mathbf{0}_{\mathbb{R}^m} \right\}, \\ (\text{Ker } \mathbf{A}^\top)^\perp &= \left\{ \mathbf{w} \in \mathbb{R}^n \mid \mathbf{w} \cdot \mathbf{y} = 0, \ \mathbf{y} \in \mathbb{R}^n, \ \mathbf{A}^\top \mathbf{y} = \mathbf{0}_{\mathbb{R}^m} \right\}. \end{aligned}$$

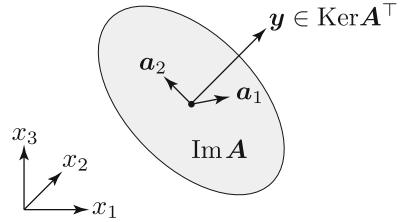
Let us choose a $\mathbf{z} \in \mathbb{R}^m$ such that, for any $\mathbf{w} \in (\text{Ker } \mathbf{A}^\top)^\perp$, $\mathbf{w} = \mathbf{A}\mathbf{z}$. Note, however, that for any $\mathbf{y} \in \text{Ker } \mathbf{A}^\top$, one can arbitrarily pick a $\mathbf{z} \in \mathbb{R}^m$ such that

$$\mathbf{w} \cdot \mathbf{y} = (\mathbf{A}\mathbf{z}) \cdot \mathbf{y} = \mathbf{z} \cdot (\mathbf{A}^\top \mathbf{y}) = 0.$$

Hence, from the definition of Eq. (A.3.2), the relation $\text{Im } \mathbf{A} = (\text{Ker } \mathbf{A}^\top)^\perp$ immediately follows. The same can be shown for $(\text{Im } \mathbf{A}^\top)^\perp = \text{Ker } \mathbf{A}$. \square

For $n = 3$ and $m = 2$, the relationship between the null space and the image space can be geometrically illustrated as in Fig. A.4.

Fig. A.4 Null space and image space



Next, let us think about the set when the equality constraint is replaced by an inequality condition. In relation to a null space, the set

$$\text{Kco } \mathbf{A} = \{ \mathbf{y} \in \mathbb{R}^m \mid \mathbf{A}\mathbf{y} \leq \mathbf{0}_{\mathbb{R}^n} \}$$

is called a non-positive cone. Moreover, with respect to an image space, the set

$$\text{Ico } \mathbf{A} = \{ \mathbf{A}\mathbf{z} \in \mathbb{R}^n \mid \mathbf{z} \geq \mathbf{0}_{\mathbb{R}^m} \}$$

is called an image cone. Furthermore, when $C \subset \mathbb{R}^n$ is a cone,

$$C' = \{ \mathbf{z} \in \mathbb{R}^n \mid \mathbf{y} \cdot \mathbf{z} \leq 0 \text{ for all } \mathbf{y} \in C \}$$

is called the dual cone of C . The next result is a rephrased version of Farkas's lemma (see, e.g., [162, Lemma 2.1, p. 27]).

Lemma A.3.2 (Farkas) *For a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$, the following relations hold:*

$$\text{Ico } \mathbf{A} = \left(\text{Kco } \mathbf{A}^\top \right)', \quad (\text{Ico } \mathbf{A})' = \text{Kco } \mathbf{A}^\top. \quad \square$$

For $n = 2$ and $m = 3$, a geometric interpretation of Farkas's lemma is shown in Fig. A.5. In this diagram,

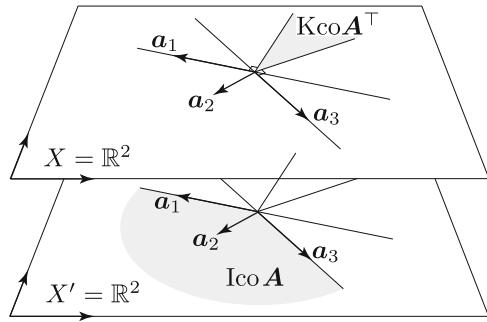
$$\text{Ico } \mathbf{A} = \left\{ (\mathbf{a}_1 \ \mathbf{a}_2 \ \mathbf{a}_3) \mathbf{z} \in \mathbb{R}^2 \mid \mathbf{z} \geq \mathbf{0}_{\mathbb{R}^3} \right\}$$

represents the vector domain (cone) in which all of \mathbf{a}_1 , \mathbf{a}_2 , \mathbf{a}_3 are in the positive direction. On the other hand,

$$\text{Kco } \mathbf{A}^\top = \left\{ \mathbf{y} \in \mathbb{R}^2 \mid \begin{pmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \mathbf{a}_3^\top \end{pmatrix} \mathbf{y} \leq \mathbf{0}_{\mathbb{R}^3} \right\}$$

represents a vector domain (cone) in which all of the inner vectors with \mathbf{a}_1 , \mathbf{a}_2 , \mathbf{a}_3 are non-positive. These mutually have the relationship of cone and dual cone.

Fig. A.5 Farkas's lemma (when $A = (a_1 \ a_2 \ a_3)$)



A.4 Implicit Function Theorem

In the proof shown in Chaps. 2 and 7 of the Lagrange multiplier method (adjoint variable method), a key result known as the implicit function theorem was applied. Here, let us recall a version of this important theorem for implicit functions on finite-dimensional vector spaces (see, e.g., [170, Section 1.37, p. 30]).

Theorem A.4.1 (Implicit Function Theorem) *Let m, n and k be natural numbers and suppose that $\mathbf{h} : \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ satisfies the following conditions in a neighborhood $B_{\mathbb{R}^m} \times B_{\mathbb{R}^n}$ of $(\mathbf{x}_0, \mathbf{y}_0) \in \mathbb{R}^m \times \mathbb{R}^n$:*

- (1) $\mathbf{h}(\mathbf{x}_0, \mathbf{y}_0) = \mathbf{0}_{\mathbb{R}^n}$,
- (2) $\mathbf{h} \in C^k(B_{\mathbb{R}^m} \times B_{\mathbb{R}^n}; \mathbb{R}^n)$,
- (3) for every $\mathbf{x} \in B_{\mathbb{R}^m}$, $\mathbf{h}(\mathbf{x}, \cdot) \in C^1(B_{\mathbb{R}^n}; \mathbb{R}^n)$, and $\mathbf{h}_{\mathbf{y}^\top}(\mathbf{x}_0, \mathbf{y}_0) = (\partial h_i / \partial y_j(\mathbf{x}_0, \mathbf{y}_0))_{ij} \in \mathbb{R}^{n \times n}$ is regular.

In this case, there exists a neighborhood $U_{\mathbb{R}^m} \times U_{\mathbb{R}^n} \subset B_{\mathbb{R}^m} \times B_{\mathbb{R}^n}$ of $(\mathbf{x}_0, \mathbf{y}_0)$ and a function $\mathbf{v} \in C^k(U_{\mathbb{R}^m}; U_{\mathbb{R}^n})$ such that $\mathbf{h}(\mathbf{x}, \mathbf{y}) = 0$ is equivalent to

$$\mathbf{y} = \mathbf{v}(\mathbf{x})$$

for all $(\mathbf{x}, \mathbf{y}) \in U_{\mathbb{R}^m} \times U_{\mathbb{R}^n}$. □

The next result is another version of implicit function theorem but for functions on Banach spaces (see, e.g., [27, Section 3.1.10, p. 115]).

Theorem A.4.2 (Implicit Function Theorem on Banach Space) *Let X, Y and Z be real Banach spaces. Suppose $h : X \times Y \rightarrow Z$ satisfies the following conditions in a neighborhood $B_X \times B_Y$ of $(\mathbf{x}_0, \mathbf{y}_0) \in X \times Y$:*

- (1) $h(\mathbf{x}_0, \mathbf{y}_0) = \mathbf{0}_Z$,
- (2) $h \in C^k(B_X \times B_Y; Z)$,
- (3) with respect to an arbitrary $\mathbf{x} \in B_X$, $h(\mathbf{x}, \cdot) \in C^1(B_Y; Z)$, and $(h_{\mathbf{y}}(\mathbf{x}, \mathbf{y}))^{-1} : Z \rightarrow Y$ is bounded and linear.

In this case, there exists a neighborhood $U_X \times U_Y \subset B_X \times B_Y$ of (x_0, y_0) and $v \in C^k(U_X; U_Y)$ such that $h(x, y) = 0$ is equivalent to

$$y = v(x)$$

for all $(x, y) \in U_X \times U_Y$. □

A.5 Lipschitz Domain

After Chap. 5, boundary value problems of partial differential equations are taken up. In defining these problems, a definition regarding the smoothness with respect to boundary of domain is used. Let us summarize these notions of regularities of domains in this section (see, e.g., [56, Definition 1.2.1.1, p. 5], [116, Definition 6.28, p. 146] for references).

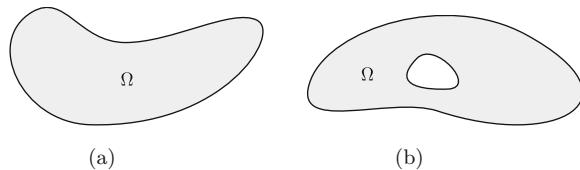
The characteristic that open sets cannot be split into two open sets (joined to be one) is called connected. For a natural number d , a connected open subset Ω in \mathbb{R}^d is called a domain. In particular, if an arbitrarily chosen closed curve in the domain can be continuously constricted to a point in the domain, as in Fig. A.6a, then the domain is called simply connected. Otherwise, the domain is called multiply connected. Figure A.6b shows an example of a doubly connected domain.

On a given domain, in order for a boundary value problem to be defined, it is not sufficient that the boundary is continuous. In fact, the additional assumption that the domain is Lipschitz is also required.

Definition A.5.1 (Lipschitz Domain) Let Ω be a $d \in \{2, 3, \dots\}$ -dimensional bounded domain with boundary denoted by $\partial\Omega$. $\partial\Omega$ is called a Lipschitz boundary if, for all $x \in \partial\Omega$, there exists a $\alpha = (\alpha_i)_i \in \mathbb{R}^d$ ($\alpha > \mathbf{0}_{\mathbb{R}^d}$) and a function (graph) φ belonging to $C^{0,1}(\mathbb{R}^{d-1}; \mathbb{R})$ (Definition 4.3.1) such that using a new coordinate system $y = (y_1, \dots, y_d)^\top = (y'^\top, y_d)^\top \in \mathbb{R}^d$ in the neighborhood

$$B(x, \alpha) = \left\{ x + (y_1, \dots, y_d)^\top \mid -\alpha_i < y_i < \alpha_i, i \in \{1, \dots, d\} \right\}$$

Fig. A.6 Domains. (a)
Simply connected domain.
(b) Doubly connected domain



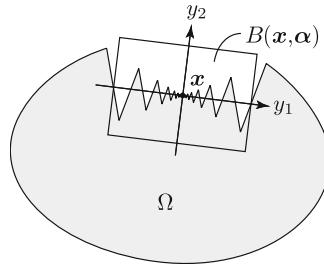


Fig. A.7 Lipschitz domain

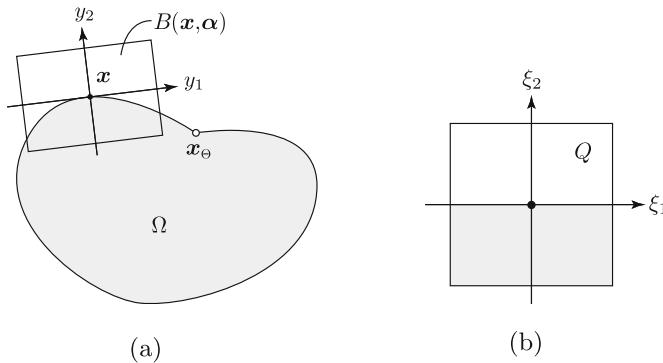


Fig. A.8 Piecewise C^1 -class boundary. (a) Neighborhood $B(x, \alpha)$ of $x \in \partial\Omega$. (b) Standard domain Q

of x , the condition

$$\Omega \cap B(x, \alpha) = \left\{ x + y = x + \left(y'^\top, y_d \right)^\top \in B(x, \alpha) \mid y_d < \varphi(y') \right\}$$

holds. Moreover, Ω is called a Lipschitz domain. \square

An extreme case of a Lipschitz domain is shown in Fig. A.7. In this figure, if the gradient of the graph φ is bounded below by some value and converges to zero as it nears x , the boundary is still Lipschitz even if it consists of regions having rapid oscillations. If the derivative of the graph is not continuous, the normal, as described later, can be defined almost everywhere on the graph except on points of discontinuity. However, for a curvature to be defined, an even smoother boundary is needed. In such situations, a mapping from the neighborhood $B(x, \alpha)$ of a point x in Fig. A.8a to a standard domain Q in Fig. A.8b can be constructed.

Definition A.5.2 (C^k -Class Domain) Let Ω be $d \in \{2, 3, \dots\}$ -dimensional bounded domain with boundary $\partial\Omega$ and $B(x, \alpha)$, as defined in Definition A.5.1, be a neighborhood of a point $x \in \partial\Omega$. For a $k \in \mathbb{N}$, $\partial\Omega$ is called a C^k -class boundary

Fig. A.9 A case where a continuous boundary cannot be defined

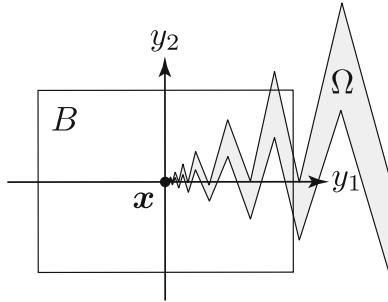
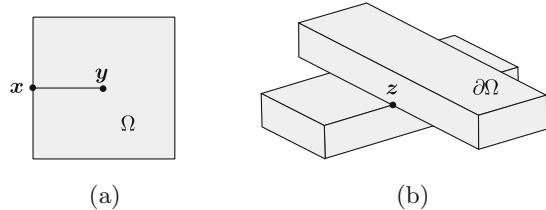


Fig. A.10 Cases where there is not a Lipschitz boundary.
(a) Domain with a crack. (b) Connected hexahedra



if there exists a function $\phi = (\phi_1, \dots, \phi_d)^\top : B(\mathbf{x}, \alpha) \rightarrow Q = (0, 1)^d$ such that

$$\Omega \cap B(\mathbf{x}, \alpha) = \{\mathbf{x} + \mathbf{y} \in B(\mathbf{x}, \alpha) \mid \phi_d(\mathbf{y}) < 0\}$$

holds for all $\mathbf{x} \in \partial\Omega$. The function ϕ and its inverse map ϕ^{-1} are uniformly bijective (one-to-one and onto mapping) and both belong to $C^k(B(\mathbf{x}, \alpha); \mathbb{R}^d)$. Moreover, Ω is called a C^k -class domain. \square

In Definition A.5.2, $k = 0$ is excluded. If a Lipschitz continuous function is chosen for ϕ in Definition A.5.2, even a domain such as that in Fig. A.9 becomes permissible [56, Section 1.2, p. 4]. Hence, an important point to note about Definition A.5.1 is that it assumes the existence of a Lipschitz continuous graph φ . Two examples of simple domains with boundaries which do not satisfy the condition of being Lipschitz are shown in Fig. A.10; refer to the neighborhood of points \mathbf{x} , \mathbf{y} and \mathbf{z} (see, e.g., [114, Figure 2, p. 91]).

Moreover, if for all $\mathbf{x} \in \Gamma$, where Γ is an open subset of $\partial\Omega$, the condition in Definition A.5.2 holds, then Γ is referred to as a C^k -class boundary. Furthermore, let the set of measure zero on the boundary where smoothness changes (see \mathbf{x}_Θ in Fig. A.8) be written as Θ . If the condition in Definition A.5.2 holds for all $\mathbf{x} \in \partial\Omega \setminus \Theta$ (points on $\partial\Omega$ excluding Θ), then $\partial\Omega$ is called a piecewise C^k -class boundary.

These definitions and notations are used to define the tangent, normal, and curvature of a point \mathbf{x} on $\partial\Omega$ as follows.

Definition A.5.3 (Tangent) Let $\partial\Omega$ be a Lipschitz boundary. For every $\mathbf{x} \in \partial\Omega \setminus \Theta$, let the neighborhood $B(\mathbf{x}, \alpha)$ and the function $\varphi \in C^{0,1}(\partial\Omega \cap B(\mathbf{x}, \alpha); \mathbb{R})$ be as

defined in Definition A.5.1. In this case, the vector

$$\begin{aligned} (\tau_1 \cdots \tau_{d-1})^\top &= \left(\frac{\partial \varphi}{\partial y_1}(\mathbf{x}) \mathbf{e}_1 \cdots \frac{\partial \varphi}{\partial y_{d-1}}(\mathbf{x}) \mathbf{e}_{d-1} \right)^\top \\ &\in \left(L^\infty(\partial\Omega \cap B(\mathbf{x}, \alpha); \mathbb{R}^d) \right)^{d-1} \end{aligned}$$

is the tangent at \mathbf{x} . Here, $\mathbf{e}_1, \dots, \mathbf{e}_{d-1}$ represent the unit vectors of the coordinate system (y_1, \dots, y_{d-1}) defined by Definition A.5.1. \square

Definition A.5.4 (Normal) Let $\partial\Omega$ be a Lipschitz boundary. Suppose $\tau_1, \dots, \tau_{d-1}$ are tangents (Definition A.5.3) at $\mathbf{x} \in \partial\Omega \setminus \Theta$. The unit vector \mathbf{v} that is orthogonal to $\tau_1, \dots, \tau_{d-1}$ and in the direction going from an inner point of Ω to the boundary is called the outer unit normal or normal at \mathbf{x} . \square

Definition A.5.5 (Mean Curvature) Suppose $\partial\Omega$ is a piecewise $C^{1,1}$ -class ($C^{1,1}$ -class on $\partial\Omega \setminus \Theta$) boundary. Given a point $\mathbf{x} \in \partial\Omega \setminus \Theta$, let $B(\mathbf{x}, \alpha)$ and the function $\phi \in C^{1,1}(B(\mathbf{x}, \alpha); \mathbb{R}^d)$ be defined as in Definition A.5.2 and let $\mathbf{v} = \nabla \phi_d(\mathbf{x}) / \|\nabla \phi_d(\mathbf{x})\|_{\mathbb{R}^d}$ and $\kappa = \nabla \cdot \mathbf{v}$. Here, $\kappa / (d-1)$ is called the mean curvature of $\partial\Omega$. \square

In what follows, we apply the definition of κ to a circle and a sphere and look at its relationship to the radius of curvature. Let us think about a circular domain with radius r centered at the origin such as that shown in Fig. A.11a. Let us calculate κ at a point $\mathbf{x} = (0, r)^\top$ on the boundary. In Fig. A.11a, the normal at the point moving from \mathbf{x} in the x_1 direction by $dx_1 = r \tan \theta$ is given by $(\sin \theta, \cos \theta)^\top$ and the normal at a point moving from \mathbf{x} in the x_2 direction by $dx_2 = y_2$ is $(0, 1)^\top$. Here,

$$\kappa(\mathbf{x}) = \nabla \cdot \mathbf{v} = \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} = \lim_{\theta \rightarrow 0} \frac{\sin \theta}{r \tan \theta} = \frac{1}{r}$$

is obtained, where r is called the radius of curvature. Moreover, at the point $\mathbf{x} = (0, 0, r)^\top$ on a sphere with a radius r centered at the origin such as that shown in Fig. A.11b, the mean curvature κ becomes twice ($d-1=2$) the value of the inverse of the radius of curvature; that is,

$$\kappa(\mathbf{x}) = \nabla \cdot \mathbf{v} = \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3} = \frac{2}{r}.$$

It should be noted that, besides the notion of mean curvature, there is also the definition for the so-called total curvature or Gauss curvature. In order to avoid confusion, let us point out the main differences between the mean and the total curvature. A smooth curved surface in a $d=3$ -dimensional space is generally a curved surface which has a spherical surface, as depicted in Fig. A.11b, replaced by an elliptical surface. At a point \mathbf{x} on the surface of such an ellipsoid, the curvatures

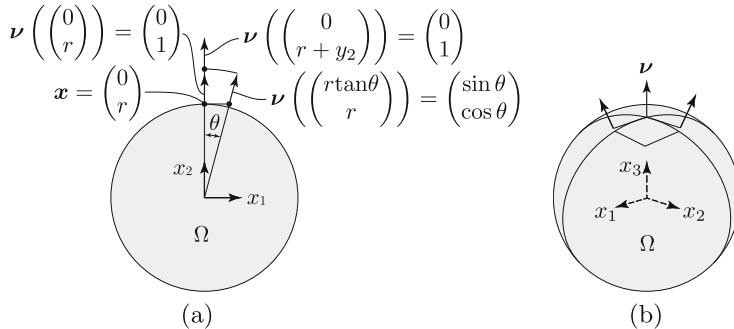


Fig. A.11 Mean curvature of circle and sphere. **(a)** $\Omega \subset \mathbb{R}^2$. **(b)** $\Omega \subset \mathbb{R}^3$

(whose sign is positive if it is concave or negative otherwise) at x of curves created on all the planes that contain the normal crossing the elliptical surface are called normal curvatures. The set of the maximum value κ_1 and minimum value κ_2 of the normal curvatures is called principal curvature. Here, the total curvature is defined by the product $\kappa_1\kappa_2$ while the mean curvature is defined by $(\kappa_1 + \kappa_2)/2$.

A.6 Heat Conduction Problem

From Chap. 5, a Poisson problem is used as a prototype problem of a boundary value problem of a partial differential equation. A Poisson problem is used as a mathematical model representing various phenomena occurring in static and balanced situations. Here, let us look at a heat conduction phenomenon as an example in order to see how a Poisson problem represents such a state of a steady heat balance. Firstly, let us consider a time-dependent heat conduction problem of a one-dimensional continuous body. Afterwards, we extend it to a time-dependent heat conduction problem of a $d \in \{2, 3\}$ -dimensional continuous body.

A.6.1 One-Dimensional Problem

Let us consider a one-dimensional continuous body such as in Fig. A.12. Let $(0, t_T)$ be a time domain and $(0, l)$ be the spatial domain for a one-dimensional continuous body. Let a be a positive real constant representing the cross sectional area. Let $b : (0, t_T) \times (0, l) \rightarrow \mathbb{R}$ be the heat released internally per unit volume and unit time, and $u : (0, t_T) \times (0, l) \rightarrow \mathbb{R}$ be the temperature distribution. Now, let us look at the steps on how to derive the equation of heat conduction for seeking u with respect to b .

Firstly, let us assume that heat and temperature are related in a certain way.

Fig. A.12 One-dimensional heat conduction problem

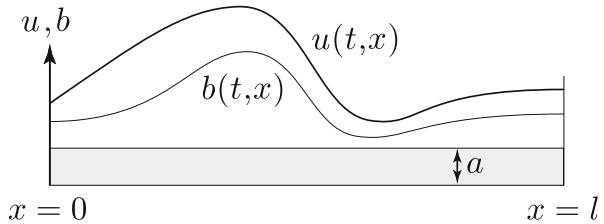
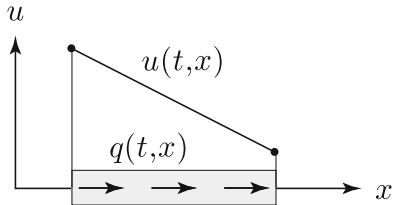


Fig. A.13 Fourier's law



Definition A.6.1 (Constitutive Equation of Heat and Temperature) Let $u(t, x)$ be the temperature distribution at $(t, x) \in (0, t_{\top}) \times (0, l)$. Then, the amount of heat per unit volume of an object to which heat conducts to is given by

$$w(t, x) = c_V(x) u(t, x), \quad (\text{A.6.3})$$

where $c_V : (0, l) \rightarrow \mathbb{R}$ is a positive-valued function representing the volume heat capacity. \square

Next, we assume that the transfer of heat follows Fourier's law of heat conduction which is given as follows (see Fig. A.13).

Definition A.6.2 (Fourier's Law of Heat Conduction) Let $u(t, x)$ be the temperature distribution at $(t, x) \in (0, t_{\top}) \times (0, l)$. The amount of heat (heat flux) passing through the cross-sectional region at x per unit time and area is given by

$$q(t, x) = -\lambda(x) \frac{\partial u}{\partial x}(t, x), \quad (\text{A.6.4})$$

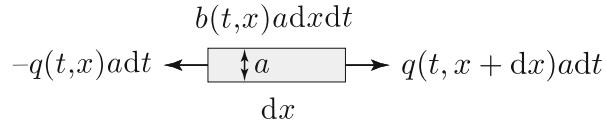
where $\lambda : (0, l) \rightarrow \mathbb{R}$ is a positive-valued function representing the heat conduction rate. \square

Here, the change in the amount of heat in an infinitesimal volume and time $adxdt$ (Fig. A.14) with respect to an arbitrary $(t, x) \in (0, t_{\top}) \times (0, l)$ becomes

$$(w(t + dt, x) - w(t, x)) dx = (b(t, x) dx - q(t, x + dx) + q(t, x)) adt.$$

In this case, if the limit of $dx \rightarrow 0, dt \rightarrow 0$ is taken,

$$\frac{\partial w}{\partial t}(t, x) = b(t, x) - \frac{\partial q}{\partial x}(t, x)$$

Fig. A.14 Heat balance

holds. Furthermore, if Eqs. (A.6.3) and (A.6.4) are used,

$$c_V \frac{\partial u}{\partial t} - \frac{\partial}{\partial x} \left(\lambda \frac{\partial u}{\partial x} \right) = b$$

is obtained. This equation is called an equation of heat conduction.

The heat conduction equation is a second-order differential equation with respect to space and a first-order differential equation with respect to time. In order to uniquely determine u , two boundary conditions and one initial condition are required. For example, the following conditions could be considered:

- (1) Let $u_D : (0, t_T) \rightarrow \mathbb{R}$ be a known temperature and suppose

$$u(t, 0) = u_D(t)$$

is satisfied at $x = 0$. This sort of condition specifying u is called a fundamental boundary condition or first-type boundary condition, and is commonly known in the literature as a Dirichlet condition.

- (2) Suppose $p_N : (0, t_T) \rightarrow \mathbb{R}$ is a known heat flux (Definition A.6.2) and that

$$\lambda \frac{\partial u}{\partial x}(t, l) = p_N(t)$$

is satisfied at $x = l$. A condition which specifies such a derivative of u is called a natural boundary condition or second-type boundary condition, and is commonly known in the literature as a Neumann condition.

- (3) Suppose $u_0 : (0, l) \rightarrow \mathbb{R}$ is a known temperature such that

$$u(0, x) = u_0(x)$$

is satisfied at $t = 0$. This sort of condition specifying u at a certain time is called an initial condition.

Initial conditions can be viewed as boundary conditions of the time domain. Hence, a problem seeking u as the solution of a partial differential equation with specified boundary conditions and initial conditions is called a boundary value problem of a partial differential equation. Heat conduction is classified as a linear second-order partial differential equation. Specifically, the heat conduction equation

can be classified as a parabolic partial differential equation (Sect. A.7). In a steady-state case, $b(t, x) = b(x)$ and $u(t, x) = u(x)$ and we get

$$-\frac{d}{dx} \left(\lambda \frac{du}{dx} \right) = b. \quad (\text{A.6.5})$$

Equation (A.6.5) is called a steady-state heat conduction equation. If this equation is extended to a d -dimensional domain, it becomes a partial differential equation such as shown later in Eq. (A.6.7). This resulting equation is classified as an elliptic partial differential equation (Sect. A.7).

A.6.2 d -Dimensional Problem

Next, let us think about a heat conduction phenomenon in a $d \in \{2, 3\}$ -dimensional object. Figure A.15 shows a two-dimensional case. Let Ω be a Lipschitz domain on \mathbb{R}^d , Γ_D be a subset of its boundary $\partial\Omega$, and $\Gamma_N = \partial\Omega \setminus \Gamma_D$. Also, let $b : (0, t_T) \times \Omega \rightarrow \mathbb{R}$ be the amount of heat emitted internally per unit time and per unit volume, and $u : (0, t_T) \times \Omega \rightarrow \mathbb{R}$ denote the temperature distribution on the domain at a given time. In this case, Fourier's law of heat conduction can be stated as follows.

Definition A.6.3 (Fourier's Law of Heat Conduction in \mathbb{R}^d) Let $u : (0, t_T) \times \Omega \rightarrow \mathbb{R}$ denote the temperature distribution on the domain at a given time. The amount of heat conducted in an object per unit time and unit area (heat flux) $\mathbf{q} : (0, t_T) \times \Omega \rightarrow \mathbb{R}^d$ satisfies

$$\mathbf{q} = \begin{pmatrix} q_1 \\ \vdots \\ q_d \end{pmatrix} = - \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1d} \\ \vdots & \ddots & \vdots \\ \lambda_{d1} & \cdots & \lambda_{dd} \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_d} \end{pmatrix} u = -\Lambda \nabla u.$$

Here, $\Lambda = (\lambda_{ij})_{ij} : \Omega \rightarrow \mathbb{R}^{d \times d}$ is a function taking the values of a positive definite real symmetric matrix (Definition 2.4.5) which represents the heat conduction rate.

Fig. A.15 Two-dimensional heat conduction problem

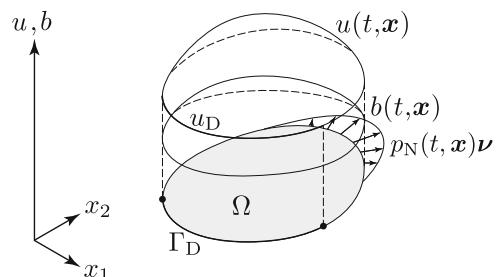
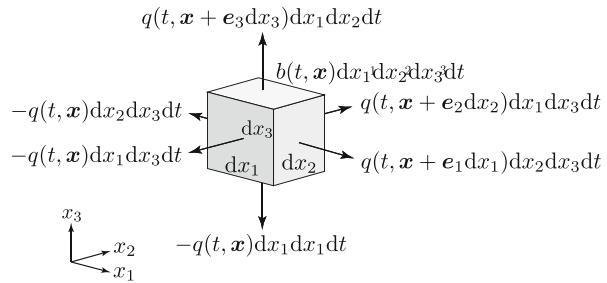


Fig. A.16 Three-dimensional heat balance



If the heat conduction rate is isotropic, a function taking positive real numbers $\lambda : \Omega \rightarrow \mathbb{R}$ can be used to write $\mathbf{A} = \lambda \mathbf{I}$ (\mathbf{I} is a unit matrix). In this case,

$$\mathbf{q} = -\lambda \nabla u. \quad (\text{A.6.6})$$

□

With respect to an arbitrary $(t, \mathbf{x}) \in (0, t_{\top}) \times \Omega$, if the change in the amount of heat in an infinitesimal $dx_1 \cdots dx_d dt$ is considered, we get

$$\begin{aligned} & (w(t + dt, \mathbf{x}) - w(t, \mathbf{x})) dx_1 dx_2 \cdots dx_d \\ &= \left\{ b(t, \mathbf{x}) - \sum_{i \in \{1, \dots, d\}} (q_i(t, \mathbf{x} + \mathbf{e}_i dx_i) - q_i(t, \mathbf{x})) \right\} dt \end{aligned}$$

where \mathbf{e}_i denotes the unit vector in the x_i -axis direction (Fig. A.16). Here, if we take the limit of $dx_1, \dots, dx_d \rightarrow 0$ and $dt \rightarrow 0$, the following equation holds:

$$\frac{\partial w}{\partial t} = b - \sum_{i \in \{1, \dots, d\}} \frac{\partial q_i}{\partial x_i}.$$

If Eqs. (A.6.3) and (A.6.6) are used,

$$\begin{aligned} & c_v \frac{\partial u}{\partial t} - \left(\frac{\partial}{\partial x_1} \cdots \frac{\partial}{\partial x_d} \right) \left(\begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1d} \\ \vdots & \ddots & \vdots \\ \lambda_{d1} & \cdots & \lambda_{dd} \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_3} \end{pmatrix} \right) u \\ &= c_v \frac{\partial u}{\partial t} - \nabla \cdot (\mathbf{A} \nabla u) = b \end{aligned}$$

is obtained. This equation is called a d -dimensional heat conduction equation. If the heat conduction rate is isotropic, we get

$$c_v \frac{\partial u}{\partial t} - \nabla \cdot (\lambda \nabla u) = b.$$

Furthermore, if λ is a real constant, we have

$$c_V \frac{\partial u}{\partial t} - \lambda \Delta u = b.$$

Here, $\Delta = \nabla \cdot \nabla$ is called the Laplace operator or harmonic operator. It can be written as ∇^2 , but in this book Δ will be used.

In order to uniquely determine the solution u which satisfies the d -dimensional heat conduction equation, certain boundary conditions have to be imposed such as the following:

- (1) Suppose $u_D : (0, t_T) \times \Gamma_D \rightarrow \mathbb{R}$ is a known temperature and that

$$u = u_D \quad \text{on } (0, t_T) \times \Gamma_D$$

is satisfied (fundamental boundary condition).

- (2) Suppose $p_N : (0, t_T) \times \Gamma_N \rightarrow \mathbb{R}$ is a known heat flux and that

$$\mathbf{v} \cdot (\mathbf{\Lambda} \nabla u) = p_N \quad \text{on } (0, t_T) \times \Gamma_N$$

is satisfied (natural boundary condition). If the heat conduction rate is isotropic, we get

$$\lambda \partial_{\mathbf{v}} u = p_N \quad \text{on } (0, t_T) \times \Gamma_N,$$

where $\partial_{\mathbf{v}}(\cdot)$ represents $(\partial(\cdot)/\partial \mathbf{x}) \cdot \mathbf{v}$.

- (3) Let $u_0 : \Omega \rightarrow \mathbb{R}$ be a known temperature and that

$$u(t_0, \mathbf{x}) = u_0(\mathbf{x}) \quad \text{in } \mathbf{x} \in \Omega$$

is satisfied with respect to $t_0 \in (0, t_T)$ (initial condition).

In a steady-state case, we have $b(t, \mathbf{x}) = b(\mathbf{x})$ and $u(t, \mathbf{x}) = u(\mathbf{x})$ and the heat conduction equation becomes

$$-\nabla \cdot (\mathbf{\Lambda} \nabla u) = b. \quad (\text{A.6.7})$$

Moreover, when a natural boundary condition is assumed across the entirety of $\partial\Omega$, there remains a constant indefiniteness. In order to uniquely determine u , the condition $|\Gamma_D| > 0$ is required (Exercise 5.2.6).

Summarizing the above, the heat conduction problem can be defined as follows. Let Ω be a $d \in \{2, 3\}$ -dimensional Lipschitz domain. Moreover, $\Gamma_D \subset \partial\Omega$ and $\Gamma_N = \partial\Omega \setminus \bar{\Gamma}_D$. Also, let $c_V : \Omega \rightarrow \mathbb{R}$ be a positive-valued function and $\mathbf{\Lambda} : \Omega \rightarrow \mathbb{R}^{d \times d}$ be a function taking positive definite real symmetric matrix values.

Problem A.6.4 (Heat Conduction Problem) Let $b : (0, t_T) \times \Omega \rightarrow \mathbb{R}$, $p_N : (0, t_T) \times \Gamma_N \rightarrow \mathbb{R}$, $u_D : (0, t_T) \times \Gamma_D \rightarrow \mathbb{R}$, $u_0 : \Omega \rightarrow \mathbb{R}$ be given. Find $u :$

$(0, t_{\top}) \times \Omega \rightarrow \mathbb{R}$ which satisfies

$$\begin{aligned} c\nu \frac{\partial u}{\partial t} - \nabla \cdot (\Lambda \nabla u) &= b && \text{in } (0, t_{\top}) \times \Omega, \\ \nu \cdot (\Lambda \nabla u) &= p_N && \text{on } (0, t_{\top}) \times \Gamma_N, \\ u &= u_D && \text{on } (0, t_{\top}) \times \Gamma_D, \\ u &= u_0 && \text{in } \Omega \text{ at } t = 0. \end{aligned} \quad \square$$

In a steady state case, the heat conduction problem can be stated as follows.

Problem A.6.5 (Steady-State Heat Conduction Problem) Let $b : \Omega \rightarrow \mathbb{R}$, $p_N : \Gamma_N \rightarrow \mathbb{R}$, $u_D : \Gamma_D \rightarrow \mathbb{R}$ be given. Find $u : \Omega \rightarrow \mathbb{R}$ such that

$$\begin{aligned} -\nabla \cdot (\Lambda \nabla u) &= b && \text{in } \Omega, \\ \nu \cdot (\Lambda \nabla u) &= p_N && \text{on } \Gamma_N, \\ u &= u_D && \text{on } \Gamma_D. \end{aligned} \quad \square$$

In Problem A.6.5, if we let $\Lambda = I$, the problem becomes a Poisson problem.

A.7 Classification of Second-Order Partial Differential Equations

As seen in Sect. A.6, a heat conduction problem is classified as parabolic if the problem is time-dependent and is classified as elliptic when viewed as a steady-state problem. Here, let us summarize the ways of classifying a system of partial differential equations (PDE) based on the standard form of a (linear) second-order partial differential equation with constant coefficients.

Definition A.7.1 (Classification of Linear Second-Order PDE) Suppose the partial differential operator $\partial/\partial x_i$ ($i \in \{1, \dots, d\}$) is expressed as ξ_i and that the characteristic equation of the term (the primary term) whose sum of orders is a maximum is $f(\xi_1, \xi_2, \dots, \xi_d) = 0$. Here, it will be classified depending on the following condition:

- (1) When the characteristic equation does not have a real solution other than $(\xi_1, \dots, \xi_d) = (0, \dots, 0)$, it is called an elliptic partial differential equation.
- (2) When the characteristic equation always has two different real solutions of $(\xi_1, \dots, \xi_d) \neq (0, \dots, 0)$, it is called a hyperbolic partial differential equation.
- (3) When the characteristic equation $f(\xi_1, \xi_2, \dots, \xi_d) = 0$ can be written as $\xi_1 - f_1(\xi_2, \dots, \xi_d) = 0$, $f_1(\xi_2, \dots, \xi_d) = 0$ has no real solution other than $(\xi_2, \dots, \xi_d) = (0, \dots, 0)$, and it is called a parabolic partial differential equation. \square

The typical form of an elliptic partial differential equation is the Laplace equation:

$$\Delta u = \left(\frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_d^2} \right) u = 0.$$

Apparently, in this case, we have

$$f(\xi_1, \dots, \xi_d) = \xi_1^2 + \cdots + \xi_d^2 = 0$$

and there are no real solutions other than $(x_1, \dots, x_d) = (0, \dots, 0)$. Apart from Laplace equations, Poisson equation $\Delta u = b$ or Helmholtz equation $\Delta u + \omega^2 u = 0$, etc., where b and ω are real numbers, are also classified as elliptic partial differential equations. These characteristics are such that they:

- are balanced-type,
- and require closed boundary conditions.

Here, “closed boundary conditions” means that at all of the points on the boundary of the domain in which a partial differential equation is defined, either a type 1 boundary condition (Dirichlet condition), type 2 boundary condition (Neumann condition) or type 3 boundary condition (Robin condition) is given. Examples include steady heat conduction (temperature), steady electric field (electric potential), steady linear elastic body (displacement), flow field of ideal gases (potential) and Stokes flow field (flow speed and pressure), etc.

On the other hand, the typical form of a hyperbolic partial differential equation is the wave motion equation:

$$\ddot{u} - c^2 \Delta u = \frac{\partial^2 u}{\partial t^2} - c^2 \left(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2} \right) u = 0,$$

where c is a positive real number and is called the wave’s speed. Consequently, in this case, the equation

$$f(\xi_1, \dots, \xi_d) = \xi_1^2 - c^2 \left(\xi_2^2 + \cdots + \xi_d^2 \right) = 0$$

always has two different real solutions of $(x_1, \dots, x_d) \neq (0, \dots, 0)$. The characteristics of a hyperbolic partial differential equation are that:

- it is time-dependent,
- and requires a closed boundary condition and two initial conditions.

Furthermore, the typical form of a parabolic partial differential equation is the diffusion equation:

$$\dot{u} - a \Delta u = \frac{\partial u}{\partial t} - a \left(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2} \right) u = 0,$$

where a is a positive real number and is called the diffusion coefficient. Accordingly, in this case, we have

$$f(\xi_1, \dots, \xi_d) = \xi_1 - a(\xi_2^2 + \dots + \xi_d^2) = 0$$

The characteristics of a parabolic partial differential equation are that:

- it is time-dependent,
- and requires closed boundary conditions and one initial condition.

A.8 Divergence Theorems

Beyond Chap. 5, integral formulae based on divergence theorems are used frequently. Here, we recall Gauss's divergence theorem and the Gauss–Green theorem.

Theorem A.8.1 (Gauss's Divergence Theorem) *Let Ω be a d -dimensional, $d \in \{2, 3, \dots\}$, Lipschitz domain and \mathbf{v} be an outward unit normal (Definition A.5.4) on boundary $\partial\Omega$. In this case, the identity*

$$\int_{\Omega} \nabla f \, dx = \int_{\partial\Omega} f \mathbf{v} \, d\gamma$$

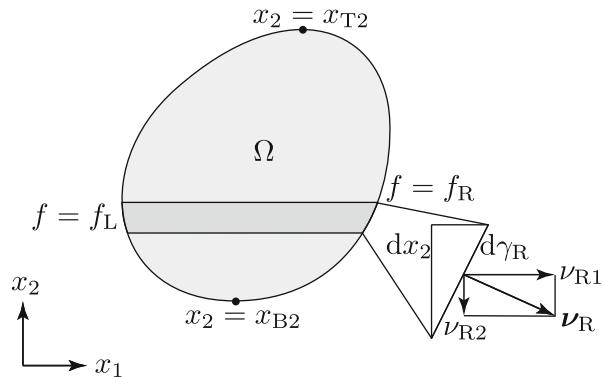
holds for any function $f \in C^1(\Omega; \mathbb{R})$. □

Proof We only outline the proof of the theorem. For more details, we refer the readers to [114, Theorem 3.34, p. 97]. Suppose $\Omega \subset \mathbb{R}^2$ is convex. The integral of $\partial f / \partial x_1$ on Ω can change the result of integrating with a very small domain in a band form of height dx_2 such as in Fig. A.17 into one, which is integrated over the interval x_2 . Here, we have

$$\int_{\Omega} \frac{\partial f}{\partial x_1} \, dx = \int_{x_{B2}}^{x_{T2}} (f_R - f_L) \, dx_2 = \int_{\partial\Omega} f \mathbf{v}_1 \, d\gamma.$$

The same results can be obtained with respect to $\partial f / \partial x_2$. Moreover, if $\Omega \subset \mathbb{R}^2$ is not convex, it can be split into convex partial domains and similar results can be obtained in the split domains. The same results can be established even when Ω is extended to $d \in \{3, 4, \dots\}$ -dimensions. □

Fig. A.17 Gauss's divergence theorem



Theorem A.8.2 (Gauss–Green Theorem) Let Ω be a d -dimensional, $d \in \{2, 3, \dots\}$, Lipschitz domain and ν be an outward unit normal on the boundary $\partial\Omega$. Also, let $p, q \in (1, \infty)$ be such that $1/p + 1/q = 1$, and assume $f \in W^{1,p}(\Omega; \mathbb{R})$ and $g \in W^{1,q}(\Omega; \mathbb{R})$. In this case, the following identity holds:

$$\int_{\Omega} \nabla f g \, dx = \int_{\partial\Omega} f g \nu \, d\gamma - \int_{\Omega} f \nabla g \, dx. \quad \square$$

Proof We only give the main working equation where the identity follows. For a more detailed proof, we refer the readers to [56, Theorem 1.5.3.1, p. 52]. From Leibnitz's law and Gauss's divergence theorem, the following equation holds:

$$\int_{\Omega} \nabla f g \, dx = \int_{\Omega} \nabla (fg) \, dx - \int_{\Omega} f \nabla g \, dx = \int_{\partial\Omega} f g \nu \, d\gamma - \int_{\Omega} f \nabla g \, dx. \quad \square$$

A.9 Inequalities

Theorems with respect to several inequalities are used in the proof of theorems and propositions beyond Chap. 4. Here, we shall summarize the inequalities used this book.

We first state the well-known Hölder's inequality (see, e.g., [105, Theorem 2.8, p. 40]) which we have used in proving that the Lebesgue space $L^p(\Omega; \mathbb{R})$ (Definition 4.3.3) is a Banach space in Chap. 4, and also in investigating the regularity of solutions of the state determination problems in Chaps. 8 and 9.

Theorem A.9.1 (Hölder's Inequality) Let Ω be a measurable set on \mathbb{R}^d , $d \in \mathbb{N}$, and $p, q \in (1, \infty)$ be such that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Then, for any pair of functions $f \in L^p(\Omega; \mathbb{R})$ and $g \in L^q(\Omega; \mathbb{R})$, the following inequality holds:

$$\|fg\|_{L^1(\Omega; \mathbb{R})} \leq \|f\|_{L^p(\Omega; \mathbb{R})} \|g\|_{L^q(\Omega; \mathbb{R})}. \quad \square$$

In Theorem A.9.1, when $p = q = 2$, it is called Schwarz's inequality.

Furthermore, in showing that a Lebesgue space $L^p(\Omega; \mathbb{R})$ is a linear space in Chap. 4, we have used Minkowski's inequality which can be stated as follows (see, e.g., [105, Theorem 2.9, p. 41]).

Theorem A.9.2 (Minkowski's Inequality) *Let Ω be a measurable set on \mathbb{R}^d , $d \in \mathbb{N}$, and suppose $p \in [1, \infty)$. Then, for any pair of functions $f, g \in L^p(\Omega; \mathbb{R})$, the following inequality holds:*

$$\|f + g\|_{L^p(\Omega; \mathbb{R})} \leq \|f\|_{L^p(\Omega; \mathbb{R})} + \|g\|_{L^p(\Omega; \mathbb{R})}. \quad \square$$

Theorem A.9.2 is equivalent to triangle inequality at $L^p(\Omega; \mathbb{R})$.

Next, we state the widely known Poincaré's inequality (see, e.g, [56, Theorem 1.4.3.4, p. 26], [41, Theorem 6.1-2 (b), p. 276]) which is a key result used in Chap. 5 when showing the existence of a unique solution to the Poisson problem (see Exercise 5.2.5). The said inequality was also used in error analyses in Chaps. 8 and 9.

Theorem A.9.3 (Poincaré Inequality) *Let Ω be a measurable set on \mathbb{R}^d , $d \in \mathbb{N}$, and suppose $p \in [1, \infty)$. Then, for a function $f \in W^{1,p}(\Omega; \mathbb{R})$, there exists a positive constant c depending only on Ω and p , such that the inequality*

$$\|f - f_0\|_{L^p(\Omega; \mathbb{R})} \leq c \|\nabla f\|_{L^p(\Omega; \mathbb{R}^d)}, \quad f_0 = \frac{1}{|\Omega|} \int_{\Omega} f \, dx$$

holds, where $|\Omega| = \int_{\Omega} dx$. Moreover, when $f = 0$ on $\Gamma_D \subset \partial\Omega$ such that $|\Gamma_D| > 0$,

$$\|f\|_{L^p(\Omega; \mathbb{R})} \leq c \|\nabla f\|_{L^p(\Omega; \mathbb{R}^d)}$$

holds. \square

Corollary A.9.4 (Poincaré Inequality) *Let Ω be a measurable set on \mathbb{R}^d , $d \in \mathbb{N}$. Also, let $k \in \mathbb{N}$ and $p \in [1, \infty)$. If $f = 0$ on $\Gamma_D \subset \partial\Omega$ such that $|\Gamma_D| > 0$, then there exists a positive constant c which depends only on Ω and p such that the inequality*

$$|f|_{W^{k,p}(\Omega; \mathbb{R})} \leq \|f\|_{W^{k,p}(\Omega; \mathbb{R})} \leq c |f|_{W^{k,p}(\Omega; \mathbb{R})}$$

holds, where $|\cdot|_{W^{k,p}(\Omega; \mathbb{R}^d)}$ represents

$$|f|_{W^{k,p}(\Omega; \mathbb{R})} = \begin{cases} \left(\sum_{|\beta|=k} \left\| \nabla^\beta f \right\|_{L^p(\Omega; \mathbb{R})}^p \right)^{1/p} & \text{for } p \in [0, \infty), \\ \max_{|\beta|=k} \left\| \nabla^\beta f \right\|_{L^\infty(\Omega; \mathbb{R})} & \text{for } p = \infty. \end{cases} \quad \square$$

Furthermore, with respect to a function of $\Omega \rightarrow \mathbb{R}^d$, such as that in the case of a linear elastic problem, Korn's inequality (for example, [165, Lemma 5.4.21, p. 312]) and Korn's second inequality shown next are used (for example, [165, Lemma 5.4.18, p. 312], where $\Gamma_D = \partial\Omega$ is assumed and $c = 2$ is obtained).

Theorem A.9.5 (Korn's Inequality) *Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain. For a function $f = (f_i)_i \in H^1(\Omega; \mathbb{R}^d)$, let*

$$\mathbf{E}(f) = (e_{ij}(f))_{ij} = \frac{1}{2} \left(\frac{\partial f_i}{\partial x_j} + \frac{\partial f_j}{\partial x_i} \right)_{ij},$$

$$\|\mathbf{E}(f)\|_{L^2(\Omega; \mathbb{R}^{d \times d})}^2 = \int_{\Omega} \sum_{(i,j) \in \{1, \dots, d\}^2} e_{ij}(f) e_{ij}(f) \, dx.$$

Then, there exist positive constants c_1 and c_2 which are only dependent on Ω such that the inequality

$$\|\mathbf{E}(f)\|_{L^2(\Omega; \mathbb{R}^{d \times d})}^2 \geq c_1 \|f\|_{H^1(\Omega; \mathbb{R}^d)}^2 - c_2 \|f\|_{L^2(\Omega; \mathbb{R}^d)}^2$$

holds. \square

Theorem A.9.6 (Korn's Second Inequality) *In Theorem A.9.5, when $f = \mathbf{0}_{\mathbb{R}^d}$ on $\Gamma_D \subset \partial\Omega$ such that $|\Gamma_D| > 0$, there exists a positive constant c depending only on Ω such that*

$$\|f\|_{H^1(\Omega; \mathbb{R}^d)}^2 \leq c \|\mathbf{E}(f)\|_{L^2(\Omega; \mathbb{R}^{d \times d})}^2$$

holds. \square

A.10 Ascoli–Arzelà Theorem

In this book, optimization problems in which continuous functions are chosen as design variables are considered in Chaps. 8 and 9. In these cases, admissible sets for design variables are required to be compact in the linear spaces where the design

variables were defined. Regarding the compactness of set of continuous functions, the following theorem is known (cf. [184, Section III.3, p. 85]).

Theorem A.10.1 (Ascoli–Arzelà Theorem) *Let X be a complete metric space and V be a compact subset of X . If a sequence $\{f_n\}_{n \in \mathbb{N}}$ of continuous functions on V is uniformly bounded and equicontinuous on V , then there exists a subsequence of $\{f_n\}_{n \in \mathbb{N}}$ that converges uniformly.* \square

This theorem assures that a bounded subset of a continuous function's set $C^k(\Omega; \mathbb{R})$ defined on a bounded domain $\Omega \subset \mathbb{R}^d$ has a subsequence converging uniformly (compact) with $\|\cdot\|_{C^k(\Omega; \mathbb{R})}$. This result shows that when considering an optimization problem constructed with continuous functions as design variables, if a bounded sequence of functions included in a bounded subset of $C^k(\Omega; \mathbb{R})$ could be found by an optimization algorithm and is confirmed to be converged, then the convergent point is in the bounded subset.

Answers to Practice Problems

Chapter 1

1.1 Let the Lagrange function of f_0 be

$$\begin{aligned}\mathcal{L}_0(\mathbf{a}, \mathbf{u}, \mathbf{v}_0) &= f_0(\mathbf{u}) + \mathcal{L}_S(\mathbf{a}, \mathbf{u}, \mathbf{v}_0) \\ &= f_0(\mathbf{u}) - \mathbf{v}_0 \cdot (\mathbf{K}(\mathbf{a})\mathbf{u} - \mathbf{p}) \\ &= (0 \ u_2) \begin{pmatrix} 0 \\ u_2 \end{pmatrix} \\ &\quad - (v_{01} \ v_{02}) \left(\frac{e_Y}{l} \begin{pmatrix} a_1 + a_2 & -a_2 \\ -a_2 & a_2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} - \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \right),\end{aligned}$$

where $\mathbf{v}_0 \in \mathbb{R}^2$ is an adjoint variable (Lagrange multiplier). The stationary condition of \mathcal{L}_0 with respect to an arbitrary variation $\hat{\mathbf{v}}_0 \in U$ of \mathbf{v}_0 ,

$$\mathcal{L}_{0\mathbf{v}_0}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\hat{\mathbf{v}}_0] = \mathcal{L}_S(\mathbf{a}, \mathbf{u}, \hat{\mathbf{v}}_0) = 0$$

holds when \mathbf{u} satisfies the state equation. The stationary condition of \mathcal{L}_0 with respect to an arbitrary variation $\hat{\mathbf{u}} \in U$ of \mathbf{u} :

$$\begin{aligned}\mathcal{L}_{0\mathbf{u}}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\hat{\mathbf{u}}] &= f_{0\mathbf{u}}(\mathbf{u})[\hat{\mathbf{u}}] - \mathcal{L}_{S\mathbf{u}}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\hat{\mathbf{u}}] \\ &= 2(0 \ u_2) \begin{pmatrix} \hat{u}_1 \\ \hat{u}_2 \end{pmatrix} - \mathbf{v}_0 \cdot (\mathbf{K}(\mathbf{a})\hat{\mathbf{u}}) \\ &= -\hat{\mathbf{u}} \cdot \left(\mathbf{K}^\top(\mathbf{a})\mathbf{v}_0 - \begin{pmatrix} 0 \\ 2u_2 \end{pmatrix} \right) = 0\end{aligned}$$

holds when v_0 satisfies

$$\frac{e_Y}{l} \begin{pmatrix} a_1 + a_2 & -a_2 \\ -a_2 & a_2 \end{pmatrix} \begin{pmatrix} v_{01} \\ v_{02} \end{pmatrix} = \begin{pmatrix} 0 \\ 2u_2 \end{pmatrix}. \quad (\text{P.1.1})$$

Equation (P.1.1) is an adjoint equation with respect to f_0 . Moreover, when u satisfies the state equation and v_0 is the solution of Eq. (P.1.1), the following, which is the same as Eq. (1.1.36), can be obtained:

$$\begin{aligned} \mathcal{L}_{0a}(\mathbf{a}, \mathbf{u}, v_0)[\mathbf{b}] &= f'_0(\mathbf{u}(\mathbf{a}))[\mathbf{b}] \\ &= - \left\{ v_0 \cdot \left(\frac{\partial \mathbf{K}(\mathbf{a})}{\partial a_1} \mathbf{u} \frac{\partial \mathbf{K}(\mathbf{a})}{\partial a_2} \mathbf{u} \right) \right\} \mathbf{b} \\ &= l(-\sigma(u_1)\varepsilon(v_{01}) - \sigma(u_2 - u_1)\varepsilon(v_{02} - v_{01})) \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \\ &= \mathbf{g}_0 \cdot \mathbf{b}. \end{aligned}$$

1.2 If u satisfies $\min_{u \in \mathbb{R}^2} \pi(\mathbf{a}, u)$,

$$\pi_u(\mathbf{a}, u)[\hat{u}] = \hat{u} \cdot (\mathbf{K}(\mathbf{a})u - \mathbf{p}) = 0$$

holds with respect to an arbitrary $\hat{u} \in \mathbb{R}^2$. In other words, it is satisfied if u is the solution of the state determination problem (Problem 1.1.3). Moreover, there exists $\alpha > 0$ such that

$$\pi_{uu}(\mathbf{a}, u)[\hat{u}, \hat{u}] = \hat{u} \cdot (\mathbf{K}(\mathbf{a})\hat{u}) > \alpha \|\hat{u}\|_{\mathbb{R}^2}^2.$$

Hence, it can be confirmed that the solution u of the state determination problem (Problem 1.1.3) is a minimizer of $\pi(\mathbf{a}, u)$. On the other hand, the maximum point of $\pi(\mathbf{a}, u)$ with respect to \mathbf{a} becomes the minimum point of $-\pi(\mathbf{a}, u)$. When u is the solution to a state determination problem,

$$\begin{aligned} -\pi_a(\mathbf{a}, u)[\mathbf{b}] &= -\frac{1}{2} \left\{ \mathbf{u} \cdot \left(\frac{\partial \mathbf{K}(\mathbf{a})}{\partial a_1} \mathbf{u} \frac{\partial \mathbf{K}(\mathbf{a})}{\partial a_2} \mathbf{u} \right) \right\} \mathbf{b} \\ &= -\frac{1}{2} \frac{e_Y}{l} (u_1 u_1 (u_2 - u_1) (u_2 - u_1)) \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \\ &= \frac{1}{2} \mathbf{g}_0 \cdot \mathbf{b} \end{aligned}$$

holds with respect to an arbitrary $\mathbf{b} \in \mathbb{R}^2$. Here, \mathbf{g}_0 expresses the vector of Eq. (1.1.36).

1.3 Since \mathbf{u} is obtained by Eq. (1.1.20),

$$f_0(\mathbf{u}(\mathbf{a})) = \left(\frac{2}{a_1} + \frac{1}{a_2} \right)^2.$$

As per Exercise 1.1.7, let

$$\tilde{f}_0(a_1) = f_0(\mathbf{u}(a_1, 1 - a_1)) = \left(\frac{2}{a_1} + \frac{1}{1 - a_1} \right)^2.$$

Here, the values of a_1 that satisfy

$$\frac{d\tilde{f}_0}{da_1} = 2 \left(\frac{2}{a_1} + \frac{1}{1 - a_1} \right) \left\{ -\frac{2}{a_1^2} + \frac{1}{(1 - a_1)^2} \right\} = 0$$

are 2 , $2 - \sqrt{2}$ and $2 + \sqrt{2}$. The values of a_2 with respect to these are -1 , $\sqrt{2} - 1$ and $-\sqrt{2} - 1$, respectively. Of these, the one satisfying $\mathbf{a} \geq \mathbf{0}_{\mathbb{R}^2}$ is determined when $\mathbf{a} = (2 - \sqrt{2}, \sqrt{2} - 1)^T$. Moreover, due to the convexity of \tilde{f}_0 and f_1 , this \mathbf{a} , which satisfies the KKT conditions, is the minimizer of Practice 1.1.

1.4 The side constraint with respect to the cross-sectional area a_1 in the definition of admissible set \mathcal{D} in Eq. (1.1.16) of design variable \mathbf{a} becomes active. Hence, in addition to $f_1(\mathbf{a}) \leq 0$, the second inequality constraint is set to be

$$f_2(\mathbf{a}) = a_{01} - a_1 \leq 0.$$

Here, the cross-sectional derivative of f_2 is

$$f_{2\mathbf{a}} = \begin{pmatrix} -1 \\ 0 \end{pmatrix} = \mathbf{g}_2. \quad (\text{P.1.2})$$

If the Lagrange multiplier with respect to $f_2 \leq 0$ is set to be λ_2 , the KKT conditions are given by

$$\mathcal{L}_{\mathbf{a}}(\mathbf{a}, \lambda_1, \lambda_2) = \mathbf{g}_0 + \lambda_1 \mathbf{g}_1 + \lambda_2 \mathbf{g}_2 = \mathbf{0}_{\mathbb{R}^2}, \quad (\text{P.1.3})$$

$$\mathcal{L}_{\lambda_1}(\mathbf{a}, \lambda_1, \lambda_2) = f_1(\mathbf{a}) = l(a_1 + a_2) - c_1 \leq 0,$$

$$\mathcal{L}_{\lambda_2}(\mathbf{a}, \lambda_1, \lambda_2) = f_2(\mathbf{a}) = a_{01} - a_1 \leq 0,$$

$$\lambda_1 f_1(\mathbf{a}) = 0,$$

$$\lambda_2 f_2(\mathbf{a}) = 0,$$

$$\lambda_1 \geq 0,$$

$$\lambda_2 \geq 0.$$

With an optimal solution, $f_1 = 0$, $f_2 = 0$, $\lambda_1 > 0$ and $\lambda_2 > 0$. Here, if \mathbf{g}_0 , \mathbf{g}_1 and \mathbf{g}_2 of Eqs. (1.1.28), (1.1.17) and (P.1.2), respectively, are substituted into Eq. (P.1.3),

$$l \begin{pmatrix} -\sigma(u_1) \varepsilon(u_1) \\ -\sigma(u_2 - u_1) \varepsilon(u_2 - u_1) \end{pmatrix} + \lambda_1 \begin{pmatrix} l \\ l \end{pmatrix} + \lambda_2 \begin{pmatrix} -1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Here, if the simultaneous linear equations with respect to λ_1 and λ_2 are solved, we obtain

$$\begin{aligned} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} &= \begin{pmatrix} \sigma(u_2 - u_1) \varepsilon(u_2 - u_1) \\ -l\sigma(u_1) \varepsilon(u_1) + l\sigma(u_2 - u_1) \varepsilon(u_2 - u_1) \end{pmatrix} \\ &= \sigma(u_2 - u_1) \varepsilon(u_2 - u_1) \begin{pmatrix} 1 \\ l \end{pmatrix}. \end{aligned}$$

1.5 Let us use the adjoint variable method. Equation (1.1.36) becomes

$$\begin{aligned} \mathcal{L}_{0a}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0) [\mathbf{b}] &= - \left\{ \mathbf{v}_0 \cdot \left(\frac{\partial \mathbf{K}(\mathbf{a})}{\partial a_1} \mathbf{u} \frac{\partial \mathbf{K}(\mathbf{a})}{\partial a_2} \mathbf{u} \right) \right\} \mathbf{b} \\ &= -\frac{e_Y}{l} (v_{01} \ v_{02}) \left(\begin{pmatrix} 2a_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \begin{pmatrix} 2a_2 & -2a_2 \\ -2a_2 & 2a_2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \right) \mathbf{b} \\ &= -\frac{e_Y}{l} (2a_1 u_1 v_{01} \ 2a_2 (u_2 - u_1) (v_{02} - v_{01})) \mathbf{b} \\ &= \mathbf{g}_0 \cdot \mathbf{b}. \end{aligned}$$

Here, if the self-adjoint relationship (Eq. (1.1.35)) is used, we get

$$\mathbf{g}_0 = -\frac{e_Y}{l} \begin{pmatrix} 2a_1 u_1^2 \\ 2a_2 (u_2 - u_1)^2 \end{pmatrix}.$$

The Hesse matrix is calculated as shown below. The second-order derivative of the Lagrange function \mathcal{L}_0 with respect to arbitrary variations $(\mathbf{b}_1, \hat{\mathbf{u}}_1)$ and $(\mathbf{b}_2, \hat{\mathbf{u}}_2)$ of the design variable (\mathbf{a}, \mathbf{u}) becomes Eq. (1.1.38). Here, \mathbf{u} and \mathbf{v}_0 are the solutions of the state determination problem (Problem 1.1.3) and adjoint problem (Problem 1.1.5) with respect to the design variable \mathbf{a} . Furthermore, $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{u}}_2$ are taken to be the variations of \mathbf{u} given that the state determination problem is satisfied with respect to arbitrary variations \mathbf{b}_1 and \mathbf{b}_2

of \mathbf{a} , respectively. That is,

$$\begin{aligned}\hat{\mathbf{u}}(\mathbf{a})[\mathbf{b}_i] &= \frac{\partial \mathbf{u}}{\partial \mathbf{a}^\top} \mathbf{b}_i = \begin{pmatrix} \frac{\partial u_1}{\partial a_1} & \frac{\partial u_1}{\partial a_2} \\ \frac{\partial u_2}{\partial a_1} & \frac{\partial u_2}{\partial a_2} \end{pmatrix} \begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} \\ &= \begin{pmatrix} -2u_1/a_1 & 0 \\ -2u_1/a_1 & -2(u_2 - u_1)/a_2 \end{pmatrix} \begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix}.\end{aligned}$$

Here, the second-order derivative of Lagrange function \mathcal{L}_0 is

$$\begin{aligned} &(\mathcal{L}_{0a}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\mathbf{b}_1] + \mathcal{L}_{0u}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\hat{\mathbf{u}}(\mathbf{a})[\mathbf{b}_1]])_{\mathbf{a}}[\mathbf{b}_2] \\ &+ (\mathcal{L}_{0a}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\mathbf{b}_1] + \mathcal{L}_{0u}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\hat{\mathbf{u}}(\mathbf{a})[\mathbf{b}_1]])_{\mathbf{u}}[\hat{\mathbf{u}}(\mathbf{a})[\mathbf{b}_2]] \\ &= \mathcal{L}_{Saa}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\mathbf{u}_1, \mathbf{u}_2] + 2\mathcal{L}_{Sau}(\mathbf{a}, \mathbf{u}, \mathbf{v}_0)[\mathbf{b}_1, \hat{\mathbf{u}}(\mathbf{a})[\mathbf{b}_2]] \\ &= \mathbf{b}_1 \cdot \left(\left(\frac{\partial \mathbf{g}_0}{\partial a_1} \frac{\partial \mathbf{g}_0}{\partial a_2} \right) \mathbf{b}_2 \right) \\ &- 2\mathbf{b}_1 \cdot \left(\begin{pmatrix} \mathbf{v}_0^\top \mathbf{K}_{a_1} \\ \mathbf{v}_0^\top \mathbf{K}_{a_2} \end{pmatrix} \begin{pmatrix} -2u_1/a_1 & 0 \\ -2u_1/a_1 & -2(u_2 - u_1)/a_2 \end{pmatrix} \mathbf{b}_2 \right) \\ &= -\frac{e_Y}{l} \mathbf{b}_1 \cdot \left(\begin{pmatrix} 2u_1 v_{01} & 0 \\ 0 & 2(u_2 - u_1)(v_{02} - v_{01}) \end{pmatrix} \mathbf{b}_2 \right) \\ &- 2\mathbf{b}_1 \cdot \left(\frac{e_Y}{l} \begin{pmatrix} 2a_1 v_{01} & 0 \\ -2a_2(v_{02} - v_{01}) & 2a_2(v_{02} - v_{01}) \end{pmatrix} \right. \\ &\quad \left. \times \begin{pmatrix} -2u_1/a_1 & 0 \\ -2u_1/a_1 & -2(u_2 - u_1)/a_2 \end{pmatrix} \mathbf{b}_2 \right) \\ &= -\frac{e_Y}{l} \mathbf{b}_1 \cdot \left(\begin{pmatrix} 2u_1 v_{01} & 0 \\ 0 & 2(u_2 - u_1)(v_{02} - v_{01}) \end{pmatrix} \mathbf{b}_2 \right) \\ &- \frac{2e_Y}{l} \mathbf{b}_1 \cdot \left(\begin{pmatrix} -4u_1 v_{01} & 0 \\ 0 & -4(u_2 - u_1)(v_{02} - v_{01}) \end{pmatrix} \mathbf{b}_2 \right) \\ &= \frac{6e_Y}{l} \mathbf{b}_1 \cdot \left(\begin{pmatrix} u_1 v_{01} & 0 \\ 0 & (u_2 - u_1)(v_{02} - v_{01}) \end{pmatrix} \mathbf{b}_2 \right).\end{aligned}$$

Hence, if the self-adjoint relationship (Eq. (1.1.35)) is used, we get

$$\mathbf{H}_0 = \frac{6e_Y}{l} \begin{pmatrix} u_1^2 & 0 \\ 0 & (u_2 - u_1)^2 \end{pmatrix}.$$

1.6 The cost function becomes

$$f(\mathbf{a}) = \frac{1}{6}a_1a_2.$$

Hence,

$$\mathbf{g}(\mathbf{a}) = \frac{1}{6} \begin{pmatrix} a_2 \\ a_1 \end{pmatrix}, \quad \mathbf{H} = \frac{1}{6} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Here, notice that the Hesse matrix \mathbf{H} is not positive definite.

1.7 The potential energy of Problem 1.2.1 is given by extending Eq. (1.1.9) as

$$\begin{aligned} \pi(\mathbf{u}) &= \int_0^l \frac{1}{2} \sigma(u) \varepsilon(u) a_1 dx + \cdots + \int_{(n-1)l}^{nl} \frac{1}{2} \sigma(u) \varepsilon(u) a_n dx \\ &\quad - \mathbf{p} \cdot \mathbf{u} \\ &= \frac{1}{2} \frac{e_Y}{l} a_1 u_1^2 + \cdots + \frac{1}{2} \frac{e_Y}{l} a_n (u_n - u_{n-1})^2 - p_1 u_1 - \cdots - p_n u_n. \end{aligned}$$

The stationary conditions of π which correspond to Eq. (1.2.1) can be written as

$$\begin{aligned} \frac{e_Y}{l} \begin{pmatrix} a_1 + a_2 & -a_2 & \cdots & 0 & 0 \\ -a_2 & a_2 + a_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{n-1} + a_n & -a_{n-1} \\ 0 & 0 & \cdots & -a_{n-1} & a_n \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-1} \\ u_n \end{pmatrix} \\ = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_{n-1} \\ p_n \end{pmatrix}. \end{aligned}$$

$\mathbf{K}(\mathbf{a})$ is the coefficient matrix of the left-hand side of this equation.

1.8 We use \mathbf{p} such that $\max_{\mathbf{p} \in \mathbb{R}^2} \pi(\mathbf{a}, \mathbf{p})$ satisfies

$$-\pi_{\mathbf{p}}(\mathbf{a}, \mathbf{p}) [\hat{\mathbf{p}}] = \hat{\mathbf{p}} \cdot (\mathbf{A}(\mathbf{a}) \mathbf{p} + \mathbf{u}) = 0$$

with respect to an arbitrary $\hat{\mathbf{p}} \in \mathbb{R}^2$. Here, if \mathbf{p} is the solution of the state determination problem (Problem 1.3.1), then it is satisfied. Moreover, there exists $\alpha > 0$ which satisfies

$$-\pi_{\mathbf{p} \mathbf{p}}(\mathbf{a}, \mathbf{p}) [\hat{\mathbf{p}}, \hat{\mathbf{p}}] = \hat{\mathbf{p}} \cdot (A(\mathbf{a}) \hat{\mathbf{p}}) > \alpha \|\hat{\mathbf{p}}\|_{\mathbb{R}^2}^2.$$

Hence, \mathbf{p} which satisfies the state determination problem can be confirmed to be the maximizer of $\pi(\mathbf{a}, \mathbf{p})$. On the other hand, when \mathbf{p} is a solution of the state determination problem,

$$\begin{aligned} \pi_{\mathbf{a}}(\mathbf{a}, \mathbf{p}) [\mathbf{b}] &= -\frac{1}{2} \left\{ \mathbf{p} \cdot \left(\frac{\partial \mathbf{A}(\mathbf{a})}{\partial a_1} \mathbf{p} \frac{\partial \mathbf{A}(\mathbf{a})}{\partial a_2} \mathbf{p} \right) \right\} \mathbf{b} \\ &= -\frac{1}{(a_0^2 + a_1^2 + a_2^2)^2} \left(a_1 \{a_0^2 p_1 + a_2^2 (p_1 - p_2)\}^2 \right. \\ &\quad \left. a_2 \{a_0^2 p_2 + a_1^2 (p_2 - p_1)\}^2 \right) \cdot \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \\ &= -\begin{pmatrix} \frac{u_1^2}{a_1} \\ \frac{u_2^2}{a_2} \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \\ &= \frac{1}{2} \mathbf{g}_0 \cdot \mathbf{b} \end{aligned}$$

holds with respect to an arbitrary $\mathbf{b} \in \mathbb{R}^2$. Here, \mathbf{g}_0 represents the vector of Eq. (1.3.19).

- 1.9** As in Fig. 1.9, let $\mathbf{l} = (l_0, l_1, l_2)^\top \in \mathbb{R}^3$ be the three lengths of a cylinder. Here, the value dividing the sum of the three cylinder volumes by π is given by

$$f(l_0, l_1, l_2) = r_0^2 l_0 + r_1^2 l_1 + r_2^2 l_2.$$

On the other hand, the geometric relationship leads to

$$\begin{aligned} h_1 &= l_1 \sin \theta_1 - \alpha_2 = 0, \\ h_2 &= l_0 - \alpha_1 + l_1 \cos \theta_1 = 0, \\ h_3 &= l_2 \sin \theta_2 - \beta_2 = 0, \\ h_4 &= l_0 - \beta_1 + l_2 \cos \theta_2 = 0. \end{aligned}$$

Using these relationships, we can write

$$f(l_0) = r_0^2 l_0 + r_1^2 \sqrt{\alpha_2^2 + (\alpha_1 - l_0)^2} + r_2^2 \sqrt{\beta_2^2 + (\beta_1 - l_0)^2}.$$

Here, the following can be obtained:

$$\begin{aligned}
 \frac{df}{dl_0} &= r_0^2 - \frac{r_1^2(\alpha_1 - l_0)}{\sqrt{\alpha_2^2 + (\alpha_1 - l_0)^2}} - \frac{r_2^2(\alpha_2 - l_0)}{\sqrt{\beta_2^2 + (\beta_1 - l_0)^2}} \\
 &= r_0^2 - \frac{r_1^2(\alpha_1 - l_0)}{l_1} - \frac{r_2^2(\alpha_2 - l_0)}{l_2} \\
 &= r_0^2 - r_1^2 \cos \theta_1 - r_2^2 \cos \theta_2 = 0.
 \end{aligned}$$

Chapter 2

2.1 The eigenvalues and eigenvectors of A are written as $\lambda_1 \leq \dots \leq \lambda_d \in \mathbb{R}$ and $\mathbf{x}_1, \dots, \mathbf{x}_d \in \mathbb{R}^d$ respectively. Here, the eigenvectors are mutually orthogonal, hence the arbitrary vector $\mathbf{x} \in \mathbb{R}^d$ can be written as

$$\mathbf{x} = \sum_{i \in \{1, \dots, d\}} \mathbf{x}_i \xi_i$$

by using $\xi = (\xi_1, \dots, \xi_d)^\top \in \mathbb{R}^d$. Here, let $\|\mathbf{x}_1\|_{\mathbb{R}^d} = \dots = \|\mathbf{x}_d\|_{\mathbb{R}^d} = 1$. Even with this, with respect to an arbitrary $\xi \in \mathbb{R}^d$, arbitrary $\mathbf{x} \in \mathbb{R}^d$ can be obtained. Here, if A is positive definite, from Theorem A.2.1, $\lambda_d \geq \dots \geq \lambda_1 > 0$. Hence, we get

$$\mathbf{x} \cdot A \mathbf{x} = \sum_{i \in \{1, \dots, d\}} \lambda_i \xi_i^2 \geq \lambda_1 \|\xi\|_{\mathbb{R}^d}^2 = \lambda_1 \|\mathbf{x}\|_{\mathbb{R}^d}^2 > 0$$

with respect to an arbitrary $\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}_{\mathbb{R}^d}\}$. Moreover, if A is a negative definite, $\lambda_d \leq \dots \leq \lambda_1 < 0$. Hence,

$$\mathbf{x} \cdot A \mathbf{x} = \sum_{i \in \{1, \dots, d\}} \lambda_i \xi_i^2 \geq \lambda_1 \|\xi\|_{\mathbb{R}^d}^2 = \lambda_1 \|\mathbf{x}\|_{\mathbb{R}^d}^2 < 0$$

is obtained with respect to an arbitrary $\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}_{\mathbb{R}^d}\}$.

2.2 From Theorem 2.5.2, the required conditions for f to take a minimum value are

$$\begin{aligned}
 \frac{\partial f}{\partial x_1} &= ax_1 + bx_2 + d = 0, \\
 \frac{\partial f}{\partial x_2} &= bx_1 + cx_2 + e = 0.
 \end{aligned}$$

These equations can be written as

$$\mathbf{g} = \begin{pmatrix} \partial f / \partial x_1 \\ \partial f / \partial x_2 \end{pmatrix} = \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} d \\ e \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The sufficient condition is shown by confirming that f is a convex function based on Theorem 2.5.6. In order to do so, the Hesse matrix needs to be shown to be positive semi-definite using Theorem 2.4.6. From

$$\frac{\partial^2 f}{\partial x_1 \partial x_1} = a, \quad \frac{\partial^2 f}{\partial x_1 \partial x_2} = b, \quad \frac{\partial^2 f}{\partial x_2 \partial x_2} = c,$$

the Hesse matrix becomes

$$\mathbf{H} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}.$$

If the positive definiteness of this matrix is shown by Sylvester's criterion (Theorem A.2.2), we get

$$a > 0, \quad \begin{vmatrix} a & b \\ b & c \end{vmatrix} = ac - b^2 > 0.$$

This relationship holds regardless of $\mathbf{x} \in \mathbb{R}^2$. Furthermore, if $\mathbf{b} = (d, c)^\top$ is used, $f(x_1, x_2)$ of this problem can be written as

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{2} (x_1 \ x_2) \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + (d \ c) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= \frac{1}{2} \mathbf{x} \cdot (\mathbf{H} \mathbf{x}) + \mathbf{b} \cdot \mathbf{x}. \end{aligned}$$

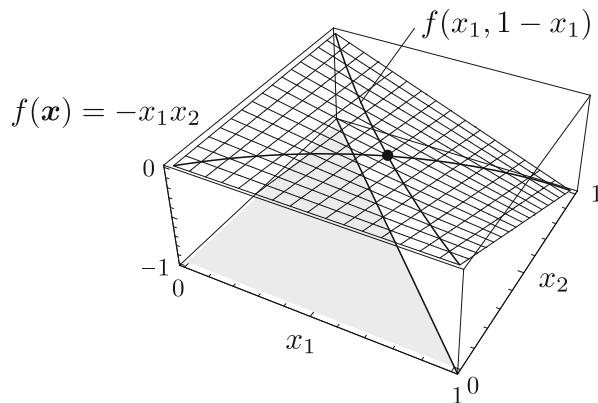
2.3 The problem can be written as

$$\min_{\mathbf{x} \in \mathbb{R}^2} \{ f_0(\mathbf{x}) = -x_1 x_2 \mid f_1(\mathbf{x}) = 2(x_1 + x_2) - c_1 \leq 0 \}.$$

Let $\lambda_1 \in \mathbb{R}$ be a Lagrange multiplier with respect to the constraint of the length of the sides of the rectangle, and the Lagrange function of this problem be

$$\mathcal{L}(x_1, x_2, \lambda_1) = f_0(\mathbf{x}) + \lambda_1 f_1(\mathbf{x}) = -x_1 x_2 + \lambda_1 \{2(x_1 + x_2) - c_1\}.$$

Fig. P.1 Function
 $f_0(\mathbf{x}) = -x_1 x_2$



The KKT conditions become

$$\mathcal{L}_{x_1} = -x_2 + 2\lambda_1 = 0,$$

$$\mathcal{L}_{x_2} = -x_1 + 2\lambda_1 = 0,$$

$$\mathcal{L}_\lambda = f_1(\mathbf{x}) = 2(x_1 + x_2) - c_1 \leq 0,$$

$$\lambda_1 f_1(\mathbf{x}) = \lambda_1 \{2(x_1 + x_2) - c_1\} = 0,$$

$$\lambda_1 \geq 0.$$

From these, the KKT conditions are satisfied when

$$\lambda_1 = \frac{x_1}{2} = \frac{x_2}{2} = \frac{c_1}{8}.$$

This result indicates a square. The fact that the solution satisfying the KKT conditions is a minimizer is shown below. f_0 is not a convex function (Exercise 2.4.9). However, $\tilde{f}_0(x_1) = f_0(x_1, -x_1 + c_1/2)$ is a convex function. Here, if it is viewed as an unconstrained minimization problem of $\tilde{f}_0(x_1)$, it can be shown that (x_1, x_2) satisfying the KKT conditions is a minimizer. Figure P.1 shows the status when $c_1 = 2$.

Chapter 3

3.1 If Eq. (3.5.7) showing the Newton–Raphson method is rewritten for f ,

$$x_{k+1} = x_k - \frac{f(x_k)}{g(x_k)}.$$

Moreover, if $g(x_k)$ is replaced by the difference,

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k)$$

is obtained.

3.2 Let $f(x_k + \bar{\epsilon}_g \bar{y}_g)$ be $\bar{f}(\bar{\epsilon}_g)$, and furthermore

$$\frac{d\bar{f}}{d\bar{\epsilon}_g}(\bar{\epsilon}_g) = \bar{g}(\bar{\epsilon}_g) = g(x_k + \bar{\epsilon}_g \bar{y}_g) \cdot \bar{y}_g.$$

In the strict line search method (Problem 3.4.1), $\bar{\epsilon}_g$ is determined so that

$$\bar{g}(\bar{\epsilon}_g) = 0$$

is satisfied. When obtaining the solution of this non-linear equation using the Newton–Raphson method, $\bar{\epsilon}_{g,l+1} = \bar{\epsilon}_{g,l} - \bar{g}(\bar{\epsilon}_{g,l}) / h(\bar{\epsilon}_{g,l})$ should be sought so that

$$\bar{g}(\bar{\epsilon}_{g,l+1}) = \bar{g}(\bar{\epsilon}_{g,l}) + h(\bar{\epsilon}_{g,l})(\bar{\epsilon}_{g,l+1} - \bar{\epsilon}_{g,l}) = 0$$

is satisfied. Here, $h(\bar{\epsilon}_{g,l})$ is a second-order derivative function of \bar{f} . When using the secant method, we would set

$$h(\bar{\epsilon}_{g,l}) = \frac{\bar{g}(\bar{\epsilon}_{g,l}) - \bar{g}(\bar{\epsilon}_{g,l-1})}{\bar{\epsilon}_{g,l} - \bar{\epsilon}_{g,l-1}}$$

and use

$$\bar{\epsilon}_{g,l+1} = \bar{\epsilon}_{g,l} - \frac{\bar{\epsilon}_{g,l} - \bar{\epsilon}_{g,l-1}}{\bar{g}(\bar{\epsilon}_{g,l}) - \bar{g}(\bar{\epsilon}_{g,l-1})} \bar{g}(\bar{\epsilon}_{g,l})$$

in order to obtain $\bar{\epsilon}_{g,l+1}$.

3.3 In the conjugate gradient method, set $x_0 = \mathbf{0}_X$ and $\bar{y}_{g,0} = -g_0 = -g(x_0) = -b$, and calculate $\bar{\epsilon}_{g,k}$ using Eq. (3.4.8) with respect to $k \in \mathbb{N} \cup \{0\}$, and x_k , g_k , β_k and $\bar{y}_{g,k}$ using from Eq. (3.4.9) to Eq. (3.4.12) with respect to $k \in \mathbb{N}$. Therefore, the following holds:

$$\begin{aligned} \bar{y}_{k+1} \cdot (\mathbf{B} \bar{y}_{g,k}) \\ = (-g_{k+1} + \beta_{k+1} \bar{y}_{g,k}) \cdot (\mathbf{B} \bar{y}_{g,k}) \\ = \left(-g_{k+1} + \frac{g_{k+1} \cdot g_{k+1}}{g_k \cdot g_k} \bar{y}_{g,k} \right) \cdot (\mathbf{B} \bar{y}_{g,k}) \end{aligned}$$

$$\begin{aligned}
&= \left\{ -\mathbf{g}_k - \bar{\epsilon}_{gk} \mathbf{B} \bar{\mathbf{y}}_{gk} + \frac{(\mathbf{g}_k + \bar{\epsilon}_{gk} \mathbf{B} \bar{\mathbf{y}}_{gk}) \cdot (\mathbf{g}_k + \bar{\epsilon}_{gk} \mathbf{B} \bar{\mathbf{y}}_{gk})}{\mathbf{g}_k \cdot \mathbf{g}_k} \bar{\mathbf{y}}_{gk} \right\} \\
&\quad \cdot (\mathbf{B} \bar{\mathbf{y}}_{gk}) \\
&= \left\{ -\mathbf{g}_k - \bar{\epsilon}_{gk} \mathbf{B} \bar{\mathbf{y}}_{gk} \right. \\
&\quad \left. + \frac{\mathbf{g}_k \cdot \mathbf{g}_k + 2\bar{\epsilon}_{gk} \mathbf{g}_k \cdot (\mathbf{B} \bar{\mathbf{y}}_{gk}) + \bar{\epsilon}_{gk}^2 (\mathbf{B} \bar{\mathbf{y}}_{gk}) \cdot (\mathbf{B} \bar{\mathbf{y}}_{gk})}{\mathbf{g}_k \cdot \mathbf{g}_k} \bar{\mathbf{y}}_{gk} \right\} \\
&\quad \cdot (\mathbf{B} \bar{\mathbf{y}}_{gk}) \\
&= \left\{ -\mathbf{g}_k - \frac{\mathbf{g}_k \cdot \mathbf{g}_k}{\bar{\mathbf{y}}_{gk} \cdot (\mathbf{B} \bar{\mathbf{y}}_{gk})} (\mathbf{B} \bar{\mathbf{y}}_{gk}) \right\} \cdot (\mathbf{B} \bar{\mathbf{y}}_{gk}) + \bar{\mathbf{y}}_{gk} \cdot (\mathbf{B} \bar{\mathbf{y}}_{gk}) \\
&\quad + 2\mathbf{g}_k \cdot (\mathbf{B} \bar{\mathbf{y}}_{gk}) + \frac{(\mathbf{B} \bar{\mathbf{y}}_{gk}) \cdot (\mathbf{B} \bar{\mathbf{y}}_{gk})}{\bar{\mathbf{y}}_{gk} \cdot (\mathbf{B} \bar{\mathbf{y}}_{gk})} \mathbf{g}_k \cdot \mathbf{g}_k \\
&= (\mathbf{g}_k + \bar{\mathbf{y}}_{gk}) \cdot (\mathbf{B} \bar{\mathbf{y}}_{gk}) = \beta_k \bar{\mathbf{y}}_{k-1} \cdot (\mathbf{B} \bar{\mathbf{y}}_{gk}).
\end{aligned}$$

Here, $\bar{\mathbf{y}}_{gk-1} \cdot \mathbf{g}_k = 0$ was used because we use the strict line search. When $k = 0$, from $\bar{\mathbf{y}}_{g0} = -\mathbf{g}_0$, $(\mathbf{g}_0 + \bar{\mathbf{y}}_{g0}) \cdot (\mathbf{B} \bar{\mathbf{y}}_{g0}) = 0$ is established. Therefore, with respect to $k \in \mathbb{N}$, $\bar{\mathbf{y}}_{gk+1} \cdot (\mathbf{B} \bar{\mathbf{y}}_{gk}) = 0$ holds.

- 3.4** The gradient $\mathbf{g}(\mathbf{a})$ and Hessian \mathbf{H} of $f(\mathbf{a})$ with respect to a variation of \mathbf{a} obtained in Practice 1.6 are used. The Newton method uses $\mathbf{H}\mathbf{b} = -\mathbf{g}(\mathbf{a}_0)$, that is,

$$\frac{1}{6} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = -\frac{1}{6} \begin{pmatrix} a_{02} \\ a_{01} \end{pmatrix}$$

to obtain the search vector \mathbf{b} . When solving this equation, we get

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = -\begin{pmatrix} a_{01} \\ a_{02} \end{pmatrix}.$$

Here, the point updated using the first Newton method:

$$\begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix} = \begin{pmatrix} a_{01} \\ a_{02} \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} a_{01} - a_{01} \\ a_{02} - a_{02} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

becomes the minimum point of f . The reason that the minimum point could be obtained after using Newton method just once is because the Taylor expansion of f is fully described by the gradient and the Hesse matrix. In that case, it can be confirmed that the positive definiteness of the Hesse matrix is not required.

Chapter 4

4.1 Let U and V be

$$U = \left\{ u \in H^1((0, l) \times (0, t_T); \mathbb{R}) \mid \begin{array}{l} u(0, t) = 0 \text{ for } t \in (0, t_T), \\ u(x, 0) = \alpha(x) \text{ for } x \in (0, l) \end{array} \right\},$$

$$V = \left\{ v \in H^1((0, l) \times (0, t_T); \mathbb{R}) \mid \begin{array}{l} v(0, t) = 0 \text{ for } t \in (0, t_T), \\ v(x, 0) = 0 \text{ for } x \in (0, l) \end{array} \right\}.$$

Select and fix an element u_0 of $H^1((0, l) \times (0, t_T); \mathbb{R})$ satisfying $u_0(x, 0) = \alpha(x)$. The first variation of $f(u)$ with respect to an arbitrary $v \in V$ becomes

$$\begin{aligned} f'(u)[v] &= \int_0^{t_T} \left\{ \int_0^l (\rho \dot{u} \dot{v} - e \nabla u \nabla v + bv) a_S dx + p_N v(l, t) a_S(l, t) \right\} dt \\ &\quad - \int_0^l \rho \beta v(x, t_T) a_S dx \\ &= \int_0^{t_T} \left\{ \int_0^l (-\rho \ddot{u} + \nabla(e \nabla u) + b) v a_S dx \right. \\ &\quad \left. - (e \nabla u(l, t) - p_N) v(l, t) a_S(l) \right\} dt \\ &\quad + \int_0^l \rho (\dot{u}(x, t_T) - \beta) v(x, t_T) a_S dx. \end{aligned}$$

Hence, the stationary condition of $f(u)$ with respect to an arbitrary $v \in V$ is given by the condition such that $f'(u)[v] = 0$ with respect to $u - u_0 \in V$. In other words, we get

$$\begin{aligned} \rho \ddot{u} - \nabla(e \nabla u) &= \rho \ddot{u} - \nabla \sigma(u) = b \text{ for } (x, t) \in (0, l) \times (0, t_T), \\ e \nabla u(l, t) &= \sigma(u(l, t)) = p_N \text{ for } t \in (0, t_T), \\ \dot{u}(x, t_T) &= \beta \text{ for } x \in (0, l). \end{aligned}$$

At that time, for $f(u)$ and $f'(u)[v]$ to have meaning, we need the following to hold:

$$\begin{aligned} \rho &\in L^\infty((0, l); \mathbb{R}), \quad \alpha \in H^1((0, l); \mathbb{R}), \quad \beta \in L^2((0, l); \mathbb{R}), \\ b &\in L^2((0, l) \times (0, t_T); \mathbb{R}), \quad p_N \in L^2((0, t_T); \mathbb{R}). \end{aligned}$$

- 4.2** The first variation of the action integral $f(\mathbf{u})$ with respect to an arbitrary variation $\mathbf{v} \in V$ of $\mathbf{u} \in U$ becomes

$$\begin{aligned}
 f'(\mathbf{u}, \dot{\mathbf{u}})[\mathbf{v}, \dot{\mathbf{v}}] &= \int_0^{t_T} \left(\frac{\partial l}{\partial \mathbf{u}} \cdot \mathbf{v} + \frac{\partial l}{\partial \dot{\mathbf{u}}} \cdot \dot{\mathbf{v}} \right) dt \\
 &= \int_0^{t_T} \left(\frac{\partial l}{\partial \mathbf{u}} - \frac{d}{dt} \frac{\partial l}{\partial \dot{\mathbf{u}}} \right) \cdot \mathbf{v} dt + \frac{\partial l}{\partial \dot{\mathbf{u}}}(t_T) \cdot \mathbf{v}(t_T) - \frac{\partial l}{\partial \dot{\mathbf{u}}}(0) \cdot \mathbf{v}(0) \\
 &= \int_0^{t_T} \left(\frac{\partial l}{\partial \mathbf{u}} - \frac{d}{dt} \frac{\partial l}{\partial \dot{\mathbf{u}}} \right) \cdot \mathbf{v} dt.
 \end{aligned}$$

With respect to an arbitrary $\mathbf{v} \in V$, for $f'(\mathbf{u}, \dot{\mathbf{u}})[\mathbf{v}, \dot{\mathbf{v}}] = 0$ to hold, the Lagrange equation of motion needs to hold.

- 4.3** The first variation of the action integral $f(\mathbf{u}, \mathbf{q})$ with respect to an arbitrary variation $\mathbf{v} \in V$ of $\mathbf{u} \in U$ and an arbitrary variation $\mathbf{r} \in Q$ of $\mathbf{q} \in Q$ becomes

$$\begin{aligned}
 f'(\mathbf{u}, \mathbf{q})[\mathbf{v}, \mathbf{r}] &= \int_0^{t_T} \left(-\dot{\mathbf{q}} \cdot \mathbf{v} - \frac{\partial \mathcal{H}}{\partial \mathbf{u}} \cdot \mathbf{v} - \dot{\mathbf{r}} \cdot \mathbf{u} - \frac{\partial \mathcal{H}}{\partial \mathbf{q}} \cdot \mathbf{r} \right) dt \\
 &= \int_0^{t_T} \left\{ -\left(\dot{\mathbf{q}} + \frac{\partial \mathcal{H}}{\partial \mathbf{u}} \right) \cdot \mathbf{v} + \left(\dot{\mathbf{u}} - \frac{\partial \mathcal{H}}{\partial \mathbf{q}} \right) \cdot \mathbf{r} \right\} dt.
 \end{aligned}$$

With respect to an arbitrary $\mathbf{v} \in V$ and an arbitrary $\mathbf{r} \in Q$, for $f'(\mathbf{u}, \mathbf{q})[\mathbf{v}, \mathbf{r}] = 0$ to hold, the Hamilton equation of motion needs to hold. Moreover, when the Hamilton equation of motion holds,

$$\dot{\mathcal{H}}(\mathbf{u}, \mathbf{q}) = \frac{\partial \mathcal{H}}{\partial \mathbf{u}} \cdot \dot{\mathbf{u}} + \frac{\partial \mathcal{H}}{\partial \mathbf{q}} \cdot \dot{\mathbf{q}} = \frac{\partial \mathcal{H}}{\partial \mathbf{u}} \cdot \frac{\partial \mathcal{H}}{\partial \mathbf{q}} - \frac{\partial \mathcal{H}}{\partial \mathbf{q}} \cdot \frac{\partial \mathcal{H}}{\partial \mathbf{u}} = 0$$

holds. Furthermore, with respect to a spring mass system of Fig. 4.1, when the external force $p = 0$, since the momentum is given by $q = m\dot{u}$, we get

$$\mathcal{H}(\mathbf{u}, q) = -l(\mathbf{u}, q) + q\dot{u} = -\frac{1}{2}m\dot{u}^2 + \frac{1}{2}ku^2 + q\dot{u} = \frac{1}{2}m\dot{u}^2 + \frac{1}{2}ku^2.$$

In other words, it shows that when there are no external forces in play, the sum of kinetic energy and potential energy becomes a Hamilton function and that it is conserved.

- 4.4** If $Y \Subset Z$, there exists some positive constant c and with respect to an arbitrary $\mathbf{x} \in Y$,

$$\|\mathbf{x}\|_Z \leq c \|\mathbf{x}\|_Y$$

holds. Here, if the definitions of norms (Definition 4.4.5) with respect to Y' and Z' are used,

$$\frac{1}{c} \|\phi\|_{Y'} = \sup_{x \in Y \setminus \{\mathbf{0}_Y\}} \frac{|\langle \phi, x \rangle|}{c \|x\|_Y} \leq \sup_{x \in Z \setminus \{\mathbf{0}_Z\}} \frac{|\langle \phi, x \rangle|}{\|x\|_Z} = \|\phi\|_{Z'}$$

is established with respect to an arbitrary $\phi \in Z'$. Therefore, from $\|\phi\|_{Y'} \leq c \|\phi\|_{Z'}, Z' \subseteq Y'$ is obtained.

Chapter 5

5.1 From the fact that Dirichlet condition is given over the whole boundary, $U = H_0^1(\Omega; \mathbb{R})$. In this case,

$$\begin{aligned} \int_{\Omega} (-\Delta u + u) v \, dx &= \int_{\Omega} (-\nabla \cdot \nabla u + u) v \, dx \\ &= - \int_{\partial\Omega} v \nabla u \cdot \mathbf{n} \, d\gamma + \int_{\Omega} (\nabla u \cdot \nabla v + uv) \, dx \\ &= \int_{\Omega} (\nabla u \cdot \nabla v + uv) \, dx = \int_{\Omega} bv \, dx \end{aligned}$$

holds with respect to an arbitrary $v \in U$. Here, the weak form of this problem becomes a problem seeking $\tilde{u} = u - u_D \in U$ satisfying

$$a(u, v) = l(v)$$

with respect to an arbitrary $v \in U$, where

$$a(u, v) = \int_{\Omega} (\nabla u \cdot \nabla v + uv) \, dx, \quad l(v) = \int_{\Omega} bv \, dx.$$

For this weak-form solution to exist uniquely, the assumptions for the Lax–Milgram theorem need to hold. $U = H_0^1(\Omega; \mathbb{R})$ is a Hilbert space. Moreover, from the fact that

$$a(v, v) = \|v\|_{H^1(\Omega; \mathbb{R})}^2$$

holds with respect to an arbitrary $v \in H_0^1(\Omega; \mathbb{R})$, a is coercive and bounded. Hence, just $\hat{l} \in U'$ needs to hold. With respect to \hat{l} ,

$$\begin{aligned} |\hat{l}(v)| &\leq \int_{\Omega} |bv| \, dx + \int_{\Omega} (|\nabla u_D \cdot \nabla v| + |u_D v|) \, dx \\ &\leq \|b\|_{L^2(\Omega; \mathbb{R})} \|v\|_{L^2(\Omega; \mathbb{R})} + \|\nabla u_D\|_{L^2(\Omega; \mathbb{R}^d)} \|\nabla v\|_{L^2(\Omega; \mathbb{R}^d)} \\ &\quad + \|u_D\|_{L^2(\Omega; \mathbb{R})} \|v\|_{L^2(\Omega; \mathbb{R})} \\ &\leq (\|b\|_{L^2(\Omega; \mathbb{R})} + \|u_D\|_{H^1(\Omega; \mathbb{R})}) \|v\|_{H^1(\Omega; \mathbb{R})} \end{aligned}$$

holds. Therefore, we need $b \in L^2(\Omega; \mathbb{R})$ and $u_D \in H^1(\Omega; \mathbb{R})$.

- 5.2** The point x_A is a boundary between a homogeneous Dirichlet and homogeneous Neumann boundaries at which the opening angle is $\alpha = \pi/2$. From Theorem 5.3.2 (2), getting $\mathbf{u} \in H^2(B_A; \mathbb{R}^2)$ around the neighborhood B_A of the point x_A , the point x_A is not a singular point. On the other hand, the point x_B is a boundary between homogeneous Neumann and non-homogeneous Neumann boundaries at which the opening angle α is $\pi/2$. There is no singularity in the solution at this angle from Theorem 5.3.2 (1). However, \mathbf{p}_N changes as a step function around the neighborhood B_B of x_B as $(0, 0)^\top$ and $(0, -1)^\top$ across the boundary Γ_p . From this, if we view it as $\mathbf{p}_N \in L^\infty(B_B; \mathbb{R}^2)$, we have $\mathbf{u} \in C^{0,1}(B_B; \mathbb{R}^2)$ which is not included in $H^2(B_B; \mathbb{R}^2)$.

- 5.3** The function space with respect to this problem is set as

$$\begin{aligned} U = \{ \mathbf{u} \in H^1((0, t_T); H^1(\Omega; \mathbb{R}^d)) \mid &\mathbf{u} = \mathbf{0}_{\mathbb{R}^d} \text{ on } \Gamma_D \times (0, t_T), \\ &\mathbf{u} = \mathbf{0}_{\mathbb{R}^d} \text{ on } \Omega \times \{0, t_T\} \}. \end{aligned}$$

Assume $\mathbf{u}_{D0}, \mathbf{u}_{DT} \in H^1(\Omega; \mathbb{R}^d)$ and $\mathbf{u}_D \in H^1((0, t_T); H^1(\Omega; \mathbb{R}^d))$. Furthermore, assume $\mathbf{b} \in L^2((0, t_T); L^2(\Omega; \mathbb{R}^d))$, $\mathbf{p}_N \in L^2((0, t_T); L^2(\Gamma_N; \mathbb{R}^d))$. Here, the weak form of this problem can be obtained by multiplying an arbitrary $\mathbf{v} \in U$ to the first equation, integrating with $\Omega \times (0, t_T)$ and using the fundamental boundary conditions as follows. “Obtain $\tilde{\mathbf{u}} = \mathbf{u} - \mathbf{u}_D \in U$ which satisfy

$$\int_0^{t_T} (b(\dot{\mathbf{u}}, \dot{\mathbf{v}}) - a(\mathbf{u}, \mathbf{v}) + l(\mathbf{v})) \, dt = 0$$

with respect to an arbitrary $\mathbf{v} \in U$, where let

$$b(\mathbf{u}, \mathbf{v}) = \int_{\Omega} \rho \mathbf{u} \cdot \mathbf{v} \, dx,$$

$$a(\mathbf{u}, \mathbf{v}) = \int_{\Omega} \mathbf{S}(\mathbf{u}) \cdot \mathbf{E}(\mathbf{v}) \, dx,$$

$$l(\mathbf{v}) = \int_{\Omega} \mathbf{b} \cdot \mathbf{v} \, dx + \int_{\Gamma_N} \mathbf{p}_N \cdot \mathbf{v} \, d\gamma.$$

5.4 Let the function space with respect to ϕ be

$$U = \left\{ \phi \in H^1(\Omega; \mathbb{R}^d) \mid \phi = \mathbf{0}_{\mathbb{R}^d} \text{ on } \Gamma_D \right\}.$$

In this case, substituting $\mathbf{u}(\mathbf{x}, t) = \phi(\mathbf{x}) e^{\lambda t}$ with respect to $\phi \in U$ into $\rho \ddot{\mathbf{u}}^\top - \nabla^\top \mathbf{S}(\mathbf{u}) = \mathbf{0}_{\mathbb{R}^d}^\top$, integrating this equation over Ω after having an arbitrary $\mathbf{v} \in U$ multiplied by it, and considering the fundamental boundary condition $\mathbf{u} = \mathbf{u}_D$ on $\Gamma_D \times (0, t_T)$, the weak form of the natural frequency problem can be obtained as below. “Obtain $(\phi, \lambda) \in U \times \mathbb{R}$ satisfying

$$\lambda^2 b(\phi, \mathbf{v}) + a(\phi, \mathbf{v}) = 0$$

with respect to an arbitrary $\mathbf{v} \in U$.”

Commentary This problem is an eigenvalue problem (the equation is an eigen equation) on a function space U . In this problem, if a non-negative definiteness (coerciveness including 0) of $a(\cdot, \cdot)$ and positive definiteness (coerciveness) of $b(\cdot, \cdot)$ are considered, eigenpairs $(\phi_i, \lambda_i)_{i \in \mathbb{N}}$ of the number of dimensions of U , which is the same as a countably infinite number, exist. In this case, $\lambda_i^2 \leq 0$, in other words, $\lambda_i = \pm i\omega_i$ (i is the imaginary unit) is derived. From this result, $\phi_i(\mathbf{x})(e^{i\omega_i t} + e^{-i\omega_i t}) = \phi_i \cos \omega_i t$ becomes a solution of the eigen value problem and ω_i and ϕ_i are called eigenfrequencies and eigenmodes.

5.5 Let the function space with respect to \mathbf{u} and p be as follows respectively:

$$U = \left\{ \mathbf{u} \in H^1((0, t_T); H^1(\Omega; \mathbb{R}^d)) \mid \mathbf{u} = \mathbf{0}_{\mathbb{R}^d} \text{ on } \partial\Omega \times (0, t_T) \cup \Omega \times \{0\} \right\},$$

$$V = \left\{ \mathbf{u} \in H^1((0, t_T); H^1(\Omega; \mathbb{R}^d)) \mid \mathbf{u} = \mathbf{0}_{\mathbb{R}^d} \text{ on } \partial\Omega \times (0, t_T) \cup \Omega \times \{t_T\} \right\},$$

$$P = \left\{ p \in L^2((0, t_T); L^2(\Omega; \mathbb{R})) \mid \int_{\Omega} p \, dx = 0 \right\}.$$

Here, if an arbitrary $\mathbf{v} \in V$ is used to multiply the Navier–Stokes equation and integrate it over $(0, t_T) \times \Omega$, and a basic boundary condition $\mathbf{u} = \mathbf{u}_D$ on $\partial\Omega \times (0, t_T) \cup \Omega \times \{0\}$ is considered, a weak-form equation with respect to the Navier–Stokes equation can be obtained. On the other hand, if an arbitrary $q \in P$ is used to multiply through the equation of continuity and integrate it over

$(0, t_T) \times \Omega$, the weak form with respect to the equation of continuity can be obtained. This can be written as below. “Obtain $(\mathbf{u} - \mathbf{u}_D, p) \in U \times Q$ ” which satisfies

$$\int_0^{t_T} (b(\dot{\mathbf{u}}, \mathbf{v}) + c(\mathbf{u})(\mathbf{u}, \mathbf{v}) + a(\mathbf{u}, \mathbf{v}) + d(\mathbf{v}, p)) dt = \int_0^{t_T} l(\mathbf{v}) dt,$$

$$\int_0^{t_T} d(\mathbf{u}, q) dt = 0$$

with respect to an arbitrary $(\mathbf{v}, q) \in U \times Q$, where we let

$$a(\mathbf{u}, \mathbf{v}) = \int_{\Omega} \mu (\nabla \mathbf{u}^T) \cdot (\nabla \mathbf{v}^T) dx,$$

$$b(\mathbf{u}, \mathbf{v}) = \int_{\Omega} \rho \mathbf{u} \cdot \mathbf{v} dx,$$

$$c(\mathbf{u})(\mathbf{w}, \mathbf{v}) = \int_{\Omega} \rho ((\mathbf{u} \cdot \nabla) \mathbf{w}) \cdot \mathbf{v} dx,$$

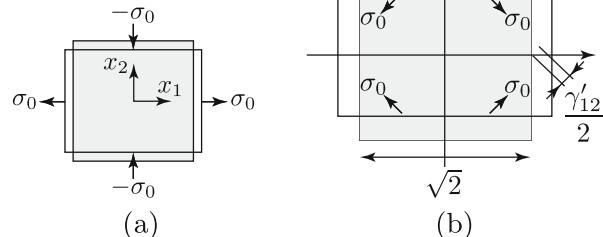
$$d(\mathbf{v}, q) = - \int_{\Omega} q \nabla \cdot \mathbf{v} dx,$$

$$l(\mathbf{v}) = \int_{\Omega} \mathbf{b} \cdot \mathbf{v} dx + \int_{\Gamma_N} \mathbf{p}_N \cdot \mathbf{v} d\gamma.$$

5.6 When a stress such as that in Fig. P.2a occurs, the linear strain becomes

$$\varepsilon_{11} = -\varepsilon_{22} = \frac{1 + \nu_p}{e_Y} \sigma_0. \quad (\text{P.5.1})$$

Fig. P.2 Deformation with shearing stress. (a) Compression and tension. (b) Coordinate system rotated in anti-clockwise direction by $\pi/4$



On the other hand, in a coordinate system which is just one $\pi/4$ rotation in the anti-clockwise direction such as in Fig. P.2b,

$$\varepsilon_{11} = \frac{\gamma'_{12}/\sqrt{2}}{\sqrt{2}} = \frac{\gamma'_{12}}{2} = \varepsilon'_{12} = \frac{\sigma_0}{2\mu_L} \quad (\text{P.5.2})$$

holds. From Eqs. (P.5.1) and (P.5.2), $e_Y = 2\mu_L (1 + \nu_P)$ holds.

Chapter 6

6.1 The weak form of this problem can be written as

$$a(u, v) + c(u, v) = l_1(v) \quad (\text{P.6.1})$$

with respect to an arbitrary $v : (0, 1) \rightarrow \mathbb{R}$ satisfying $v(0) = v(1) = 0$, where $a(\cdot, \cdot)$ and $l_1(\cdot)$ use the definitions in Exercise 6.1.5. Moreover, let

$$c(u, v) = \int_0^1 uv \, dx.$$

The result when approximate functions u_h and v_h are substituted in $a(u, v)$ and $l_1(v)$ is as per Exercise 6.1.5. Here, if u_h and v_h are substituted in $c(u, v)$, we get

$$\begin{aligned} c(u_h, v_h) &= \int_0^1 \left\{ \sum_{i \in \{1, \dots, m\}} \alpha_i \sin(i\pi x) \right\} \left\{ \sum_{j \in \{1, \dots, m\}} \beta_j \sin(j\pi x) \right\} \, dx \\ &= \boldsymbol{\beta}^\top \mathbf{C} \boldsymbol{\alpha}. \end{aligned}$$

Here, $\mathbf{C} = (c(\sin(i\pi x), \sin(j\pi x)))_{ij}$ and

$$\begin{aligned} c(\sin(i\pi x), \sin(j\pi x)) &= \int_0^1 \sin(i\pi x) \sin(j\pi x) \, dx \\ &= -\frac{1}{2} \int_0^1 [\cos((i+j)\pi x) - \cos((i-j)\pi x)] \, dx = \frac{1}{2} \delta_{ij}. \end{aligned}$$

From the answer to Exercise 6.1.5 and the result above, Eq. (P.6.1) becomes

$$(\mathbf{A} + \mathbf{C}) \boldsymbol{\alpha} = \mathbf{f}.$$

In other words,

$$\begin{aligned} & \left(\frac{\pi^2}{2} \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 4 & 0 & \cdots & 0 \\ 0 & 0 & 9 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & m^2 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \right) \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_m \end{pmatrix} \\ &= \frac{1}{\pi} \begin{pmatrix} 2 \\ 0 \\ 2/3 \\ \vdots \\ \{(-1)^{m+1} + 1\} / m \end{pmatrix}, \end{aligned}$$

or

$$\frac{i^2\pi^2 + 1}{2} \alpha_i = \frac{(-1)^{i+1} + 1}{i\pi}.$$

If this simultaneous linear equation is solved,

$$\alpha_i = \frac{2 \{(-1)^{i+1} + 1\}}{i\pi (i^2\pi^2 + 1)}$$

is obtained. Therefore, the approximate function becomes

$$u_h = \sum_{i \in \{1, \dots, m\}} \frac{2 \{(-1)^{i+1} + 1\}}{i\pi (i^2\pi^2 + 1)} \sin(i\pi x).$$

6.2 The weak form of this problem is given by Eq. (P.6.1). $a(u, v)$ and $l_1(v)$ with approximate functions u_h and v_h substituted in are as shown in Exercise 6.2.1. Here, if u_h and v_h are substituted in $c(u, v)$, we get

$$c(u_h, v_h) = \sum_{i \in \{1, \dots, m\}} \int_{x_{i-1}}^{x_i} u_h v_h \, dx = \sum_{i \in \{1, \dots, m\}} c_i(u_h, v_h),$$

$$\begin{aligned} c_i(u_h, v_h) &= (v_{i(1)} \ v_{i(2)}) \begin{pmatrix} \int_{x_{i-1}}^{x_i} \varphi_{i(1)} \varphi_{i(1)} \, dx & \int_{x_{i-1}}^{x_i} \varphi_{i(1)} \varphi_{i(2)} \, dx \\ \int_{x_{i-1}}^{x_i} \varphi_{i(2)} \varphi_{i(1)} \, dx & \int_{x_{i-1}}^{x_i} \varphi_{i(2)} \varphi_{i(2)} \, dx \end{pmatrix} \begin{pmatrix} u_{i(1)} \\ u_{i(2)} \end{pmatrix} \\ &= \bar{v}_i \cdot \bar{C}_i \bar{u}_i = \bar{v} \cdot \bar{Z}_i^\top \bar{C}_i \bar{Z}_i \bar{u} = \bar{v} \cdot \bar{C}_i \bar{u}. \end{aligned}$$

Here, $\bar{\mathbf{C}}_i = (\bar{c}_{i\alpha\beta})_{\alpha,\beta} \in \mathbb{R}^2$ becomes

$$\begin{aligned}\bar{c}_{i11} &= \int_{x_{i-1}}^{x_i} \varphi_{i(1)} \varphi_{i(1)} \, dx = \frac{1}{(x_i - x_{i-1})^2} \int_{x_{i-1}}^{x_i} (x_i - x)^2 \, dx \\ &= \frac{x_i - x_{i-1}}{3}, \\ \bar{c}_{i12} = \bar{c}_{i21} &= \int_{x_{i-1}}^{x_i} \varphi_{i(1)} \varphi_{i(2)} \, dx \\ &= \frac{1}{(x_i - x_{i-1})^2} \int_{x_{i-1}}^{x_i} (x_i - x)(x - x_{i-1}) \, dx = \frac{x_i - x_{i-1}}{6}, \\ \bar{c}_{i22} &= \int_{x_{i-1}}^{x_i} \varphi_{i(2)} \varphi_{i(2)} \, dx = \frac{1}{(x_i - x_{i-1})^2} \int_{x_{i-1}}^{x_i} (x - x_{i-1})^2 \, dx \\ &= \frac{x_i - x_{i-1}}{3}.\end{aligned}$$

In other words,

$$\bar{\mathbf{C}}_i = \frac{x_i - x_{i-1}}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Matrix $\bar{\mathbf{C}}$, which is the sum of all elements, becomes

$$\bar{\mathbf{C}} = \frac{h}{6} \begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix}.$$

Therefore, the approximate equation becomes

$$\left(\frac{1}{h} \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} + \frac{h}{6} \begin{pmatrix} 4 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 4 \end{pmatrix} \right) \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = h \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Supplementary The integrals on the finite element are simplified if the domain is changed to a standard domain. Let the mapping $\xi : (x_{i-1}, x_i) \rightarrow (0, 1)$ be

$$\xi = \frac{x - x_{i-1}}{h},$$

where $h = x_i - x_{i-1}$. Here, the Jacobian becomes

$$\frac{d\xi}{dx} = h.$$

The base function becomes

$$\varphi_{i(1)}(x) = \frac{x_i - x}{h} = 1 - \xi = \hat{\varphi}_{i(1)}(\xi),$$

$$\varphi_{i(2)}(x) = \frac{x - x_{i-1}}{h} = \xi = \hat{\varphi}_{i(2)}(\xi).$$

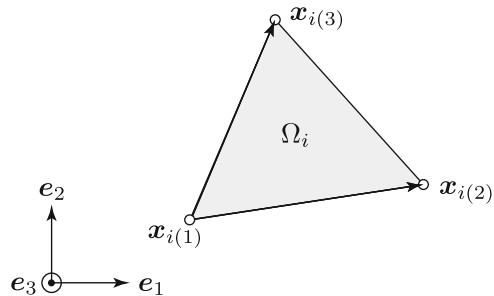
This time, \bar{C}_i can be calculated as

$$\begin{aligned}\bar{c}_{i11} &= \int_0^1 \hat{\varphi}_{i(1)} \hat{\varphi}_{i(1)} h \, d\xi = h \int_0^1 (1 - \xi)^2 \, d\xi = h \int_0^1 \eta^2 \, d\eta = \frac{h}{3}, \\ \bar{c}_{i12} = \bar{c}_{i21} &= \int_0^1 \hat{\varphi}_{i(1)} \hat{\varphi}_{i(2)} h \, d\xi = h \int_0^1 (1 - \xi)\xi \, d\xi = \frac{h}{6}, \\ \bar{c}_{i22} &= \int_0^1 \hat{\varphi}_{i(2)} \hat{\varphi}_{i(2)} h \, d\xi = h \int_0^1 \xi^2 \, d\xi = \frac{h}{3}.\end{aligned}$$

- 6.3** Let us think about a domain Ω_i of a triangular finite element such as in Fig. P.3. Here, with respect to the cross product of two vectors $\mathbf{x}_{i(2)} - \mathbf{x}_{i(1)}$ and $\mathbf{x}_{i(3)} - \mathbf{x}_{i(1)}$,

$$\begin{aligned}2 |\Omega_i| \mathbf{e}_3 \\ &= \begin{pmatrix} x_{i(2)1} - x_{i(1)1} \\ x_{i(2)2} - x_{i(1)2} \\ 0 \end{pmatrix} \times \begin{pmatrix} x_{i(3)1} - x_{i(1)1} \\ x_{i(3)2} - x_{i(1)2} \\ 0 \end{pmatrix} \\ &= \begin{vmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ x_{i(2)1} - x_{i(1)1} & x_{i(2)2} - x_{i(1)2} & 0 \\ x_{i(3)1} - x_{i(1)1} & x_{i(3)2} - x_{i(1)2} & 0 \end{vmatrix} \\ &= \begin{vmatrix} 0 & 0 & 1 \\ x_{i(2)1} - x_{i(1)1} & x_{i(2)2} - x_{i(1)2} & 0 \\ x_{i(3)1} - x_{i(1)1} & x_{i(3)2} - x_{i(1)2} & 0 \end{vmatrix} \mathbf{e}_3\end{aligned}$$

Fig. P.3 Triangular Ω_i and points $\mathbf{x}_{i(1)}$, $\mathbf{x}_{i(2)}$ and $\mathbf{x}_{i(3)}$



$$\begin{aligned}
 &= \left(\begin{vmatrix} 0 & 0 & 1 \\ x_{i(2)1} - x_{i(1)1} & x_{i(2)2} - x_{i(1)2} & 0 \\ x_{i(3)1} - x_{i(1)1} & x_{i(3)2} - x_{i(1)2} & 0 \end{vmatrix} + \begin{vmatrix} x_{i(1)1} & x_{i(1)2} & 0 \\ x_{i(1)1} & x_{i(1)2} & 1 \\ x_{i(1)1} & x_{i(1)2} & 1 \end{vmatrix} \right) \mathbf{e}_3 \\
 &= \begin{vmatrix} x_{i(1)1} & x_{i(1)2} & 1 \\ x_{i(2)1} & x_{i(2)2} & 1 \\ x_{i(3)1} & x_{i(3)2} & 1 \end{vmatrix} \mathbf{e}_3 = \gamma \mathbf{e}_3
 \end{aligned}$$

holds, where \mathbf{e}_1 , \mathbf{e}_2 and \mathbf{e}_3 are unit orthogonal vectors of x_1 , x_2 and x_3 coordinate systems. Hence, $\gamma = 2|\Omega_i|$ is obtained.

- 6.4** Let the finite elements with finite element numbers {3, 5}, {4, 6}, {1, 7} and {2, 8} be called Type 1, Type 2, Type 3 and Type 4, respectively. The result from Exercise 6.3.2 is used with respect to Type 1 and Type 2. With respect to Type 3, $\gamma = h^2$, $|\Omega_i| = h^2/2$ and

$$\begin{aligned}
 \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix} &= \frac{1}{\gamma} \begin{pmatrix} x_{i(2)2} - x_{i(3)2} \\ x_{i(3)2} - x_{i(1)2} \\ x_{i(1)2} - x_{i(2)2} \end{pmatrix} = \frac{1}{h^2} \begin{pmatrix} -h \\ h \\ 0 \end{pmatrix}, \\
 \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} &= \frac{1}{\gamma} \begin{pmatrix} x_{i(3)1} - x_{i(2)1} \\ x_{i(1)1} - x_{i(3)1} \\ x_{i(2)1} - x_{i(1)1} \end{pmatrix} = \frac{1}{h^2} \begin{pmatrix} -h \\ 0 \\ h \end{pmatrix}.
 \end{aligned}$$

Therefore,

$$\bar{\mathbf{A}}_1 = \frac{1}{2} \begin{pmatrix} 2 & -1 & -1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}, \quad \bar{\mathbf{b}}_1 = \frac{h^2}{6} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

is obtained. With respect to Type 4 too, in a similar way, $\gamma = h^2$, $|\Omega_i| = h^2/2$ and

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix} = \frac{1}{h^2} \begin{pmatrix} 0 \\ h \\ -h \end{pmatrix}, \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} = \frac{1}{h^2} \begin{pmatrix} -h \\ h \\ 0 \end{pmatrix},$$

$$\bar{A}_2 = \frac{1}{2} \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}, \quad \bar{b}_2 = \frac{h^2}{6} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

can be obtained. On the other hand, the local node number and total node numbers can be made correspondent in the way shown in Table P.1.

If a sum of all elements is taken, \bar{A} and \bar{b} become

$$\bar{A} = \frac{1}{2} \begin{pmatrix} 2 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -2 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 4 & -2 & 0 & -1 & 0 & 0 \\ 0 & -2 & 0 & -2 & 8 & -2 & 0 & -2 & 0 \\ 0 & 0 & -1 & 0 & -2 & 4 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & -2 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 2 \end{pmatrix},$$

$$\bar{b} = \frac{h^2}{6} \begin{pmatrix} 1 \\ 4 \\ 1 \\ 4 \\ 4 \\ 4 \\ 1 \\ 4 \\ 1 \end{pmatrix}.$$

Table P.1 The relationship between the local nodes $\mathbf{x}_{i(1)}$, $\mathbf{x}_{i(2)}$, $\mathbf{x}_{i(3)}$ and total nodes \mathbf{x}_j

$i \in \mathcal{E}$	1	2	3	4	5	6	7	8
$\mathbf{x}_{i(1)}$	\mathbf{x}_1	\mathbf{x}_4	\mathbf{x}_2	\mathbf{x}_2	\mathbf{x}_4	\mathbf{x}_4	\mathbf{x}_5	\mathbf{x}_8
$\mathbf{x}_{i(2)}$	\mathbf{x}_4	\mathbf{x}_5	\mathbf{x}_5	\mathbf{x}_6	\mathbf{x}_7	\mathbf{x}_8	\mathbf{x}_8	\mathbf{x}_9
$\mathbf{x}_{i(3)}$	\mathbf{x}_2	\mathbf{x}_2	\mathbf{x}_6	\mathbf{x}_3	\mathbf{x}_8	\mathbf{x}_5	\mathbf{x}_6	\mathbf{x}_6
Type	3	4	1	2	1	2	3	4

Here, the fundamental boundary conditions $u_1 = u_2 = u_3 = u_4 = u_7 = 0$ and $v_1 = v_2 = v_3 = v_4 = v_7 = 0$ and $h = 1/2$ can be used to obtain

$$\begin{pmatrix} 8 & -2 & -2 & 0 \\ -2 & 4 & 0 & -1 \\ -2 & 0 & 4 & -1 \\ 0 & -1 & -1 & 2 \end{pmatrix} \begin{pmatrix} u_5 \\ u_6 \\ u_8 \\ u_9 \end{pmatrix} = \frac{1}{12} \begin{pmatrix} 4 \\ 4 \\ 4 \\ 1 \end{pmatrix}.$$

Solving this, we get

$$\begin{pmatrix} u_5 \\ u_6 \\ u_8 \\ u_9 \end{pmatrix} = \frac{1}{16 \times 12} \begin{pmatrix} 3 & 2 & 2 & 2 \\ 2 & 6 & 2 & 4 \\ 2 & 2 & 6 & 4 \\ 2 & 4 & 4 & 12 \end{pmatrix} \begin{pmatrix} 4 \\ 4 \\ 4 \\ 1 \end{pmatrix} = \frac{1}{96} \begin{pmatrix} 15 \\ 22 \\ 22 \\ 26 \end{pmatrix}.$$

6.5 With respect to a finite element $i \in \mathcal{E}$ in Fig. 6.26, a standard domain is set to be $\Xi_i = (0, 1)^2$. The isoparametric representations of the approximate functions and coordinates become

$$\hat{u}_h(\xi) = \sum_{\alpha \in \{1, \dots, 4\}} \hat{\varphi}_\alpha(\xi) u_{i\alpha} = \hat{\varphi}(\xi) \cdot \bar{u}_i,$$

$$\hat{v}_h(\xi) = \sum_{\alpha \in \{1, \dots, 4\}} \hat{\varphi}_\alpha(\xi) v_{i\alpha} = \hat{\varphi}(\xi) \cdot \bar{v}_i,$$

$$\hat{x}_{h1}(\xi) = \sum_{\alpha \in \{1, \dots, 4\}} \hat{\varphi}_\alpha(\xi) x_{i1\alpha} = \hat{\varphi}(\xi) \cdot \bar{x}_{i1},$$

$$\hat{x}_{h2}(\xi) = \sum_{\alpha \in \{1, \dots, 4\}} \hat{\varphi}_\alpha(\xi) x_{i2\alpha} = \hat{\varphi}(\xi) \cdot \bar{x}_{i2}.$$

Here, let $x_{i1(2)} - x_{i1(1)} = h_1$ and $x_{i2(2)} - x_{i2(1)} = h_2$ and

$$\begin{pmatrix} \lambda_{11} \\ \lambda_{12} \\ \lambda_{21} \\ \lambda_{22} \end{pmatrix} = \begin{pmatrix} (x_{i1(2)} - x_1) / h_1 \\ (x_1 - x_{i1(1)}) / h_2 \\ (x_{i2(2)} - x_2) / h_1 \\ (x_2 - x_{i2(1)}) / h_2 \end{pmatrix} = \begin{pmatrix} (1 - \xi_1) \\ \xi_1 \\ (1 - \xi_2) \\ \xi_2 \end{pmatrix},$$

$$\hat{\varphi} = \begin{pmatrix} \hat{\varphi}_1(\xi) \\ \hat{\varphi}_2(\xi) \\ \hat{\varphi}_3(\xi) \\ \hat{\varphi}_4(\xi) \end{pmatrix} = \begin{pmatrix} (1 - \xi_1)(1 - \xi_2) \\ \xi_1(1 - \xi_2) \\ \xi_1 \xi_2 \\ (1 - \xi_1) \xi_2 \end{pmatrix}.$$

In this case,

$$\begin{aligned}\partial_{\xi} \hat{\varphi}_{\alpha}(\xi) &= \begin{pmatrix} \partial \hat{\varphi}_{\alpha} / \partial \xi_1 \\ \partial \hat{\varphi}_{\alpha} / \partial \xi_2 \end{pmatrix} = \begin{pmatrix} \partial \hat{x}_1 / \partial \xi_1 & \partial \hat{x}_2 / \partial \xi_1 \\ \partial \hat{x}_1 / \partial \xi_2 & \partial \hat{x}_2 / \partial \xi_2 \end{pmatrix} \begin{pmatrix} \partial \hat{\varphi}_{\alpha} / \partial x_1 \\ \partial \hat{\varphi}_{\alpha} / \partial x_2 \end{pmatrix} \\ &= \begin{pmatrix} h_1 & 0 \\ 0 & h_2 \end{pmatrix} \begin{pmatrix} \partial \hat{\varphi}_{\alpha} / \partial x_1 \\ \partial \hat{\varphi}_{\alpha} / \partial x_2 \end{pmatrix}\end{aligned}$$

holds. Hence,

$$\begin{aligned}\begin{pmatrix} \partial \hat{\varphi}_{\alpha} / \partial x_1 \\ \partial \hat{\varphi}_{\alpha} / \partial x_2 \end{pmatrix} &= \frac{1}{\omega(\xi)} \begin{pmatrix} \partial \hat{x}_2 / \partial \xi_2 & -\partial \hat{x}_2 / \partial \xi_1 \\ -\partial \hat{x}_1 / \partial \xi_2 & \partial \hat{x}_1 / \partial \xi_1 \end{pmatrix} \begin{pmatrix} \partial \hat{\varphi}_{\alpha} / \partial \xi_1 \\ \partial \hat{\varphi}_{\alpha} / \partial \xi_2 \end{pmatrix} \\ &= \frac{1}{h_1 h_2} \begin{pmatrix} h_2 & 0 \\ 0 & h_1 \end{pmatrix} \begin{pmatrix} \partial \hat{\varphi}_{\alpha} / \partial \xi_1 \\ \partial \hat{\varphi}_{\alpha} / \partial \xi_2 \end{pmatrix}\end{aligned}$$

can be obtained, where

$$\begin{aligned}\begin{pmatrix} \partial \hat{\varphi}_1 / \partial \xi_1 & \partial \hat{\varphi}_2 / \partial \xi_1 & \partial \hat{\varphi}_3 / \partial \xi_1 & \partial \hat{\varphi}_4 / \partial \xi_1 \\ \partial \hat{\varphi}_1 / \partial \xi_2 & \partial \hat{\varphi}_2 / \partial \xi_2 & \partial \hat{\varphi}_3 / \partial \xi_2 & \partial \hat{\varphi}_4 / \partial \xi_2 \end{pmatrix} \\ = \begin{pmatrix} -(1 - \xi_2) & (1 - \xi_2) \xi_2 & -\xi_2 \\ -(1 - \xi_1) & -\xi_1 & \xi_1 (1 - \xi_1) \end{pmatrix}.\end{aligned}$$

Using this result, the element coefficient matrix $\bar{A}_i = (\bar{a}_{i\alpha\beta})_{\alpha\beta} \in \mathbb{R}^{4 \times 4}$ becomes

$$\begin{aligned}\bar{a}_{i\alpha\beta} &= \int_{\Omega_i} \begin{pmatrix} \partial \varphi_{\alpha} / \partial x_1 \\ \partial \varphi_{\alpha} / \partial x_2 \end{pmatrix} \cdot \begin{pmatrix} \partial \varphi_{\beta} / \partial x_1 \\ \partial \varphi_{\beta} / \partial x_2 \end{pmatrix} dx \\ &= \int_{\Xi_i} \begin{pmatrix} \partial \hat{\varphi}_{\alpha} / \partial x_1 \\ \partial \hat{\varphi}_{\alpha} / \partial x_2 \end{pmatrix} \cdot \begin{pmatrix} \partial \hat{\varphi}_{\beta} / \partial x_1 \\ \partial \hat{\varphi}_{\beta} / \partial x_2 \end{pmatrix} \omega(\xi) d\xi \\ &= \frac{1}{h_1 h_2} \int_{\Xi_i} \begin{pmatrix} \partial \hat{\varphi}_{\alpha} / \partial \xi_1 & \partial \hat{\varphi}_{\alpha} / \partial \xi_2 \end{pmatrix} \begin{pmatrix} h_2 & 0 \\ 0 & h_1 \end{pmatrix} \begin{pmatrix} h_2 & 0 \\ 0 & h_1 \end{pmatrix} \begin{pmatrix} \partial \hat{\varphi}_{\beta} / \partial \xi_1 \\ \partial \hat{\varphi}_{\beta} / \partial \xi_2 \end{pmatrix} d\xi \\ &= \int_{\Xi_i} \left(\frac{h_2}{h_1} \frac{\partial \hat{\varphi}_{\alpha}}{\partial \xi_1} \frac{\partial \hat{\varphi}_{\beta}}{\partial \xi_1} + \frac{h_1}{h_2} \frac{\partial \hat{\varphi}_{\alpha}}{\partial \xi_2} \frac{\partial \hat{\varphi}_{\beta}}{\partial \xi_2} \right) d\xi.\end{aligned}$$

Letting $\sigma = h_2/h_1$, we get

$$\bar{a}_{i11} = \int_{\Xi_i} \left[\sigma \{-(1 - \xi_2)\}^2 + \sigma^{-1} \{-(1 - \xi_1)\}^2 \right] d\xi = \frac{1}{3} (\sigma + \sigma^{-1}).$$

From these calculations we get

$$\bar{A}_i = \frac{1}{6} \begin{pmatrix} 2\sigma + 2\sigma^{-1} & -2\sigma + \sigma^{-1} & -\sigma - \sigma^{-1} & \sigma - 2\sigma^{-1} \\ -2\sigma + \sigma^{-1} & 2\sigma + 2\sigma^{-1} & \sigma - 2\sigma^{-1} & -\sigma - \sigma^{-1} \\ -\sigma - \sigma^{-1} & \sigma - 2\sigma^{-1} & 2\sigma + 2\sigma^{-1} & -2\sigma + \sigma^{-1} \\ \sigma - 2\sigma^{-1} & -\sigma - \sigma^{-1} & -2\sigma + \sigma^{-1} & 2\sigma + 2\sigma^{-1} \end{pmatrix}.$$

The known term vector $\bar{l}_i = (\bar{l}_{i\alpha})_\alpha \in \mathbb{R}^4$ becomes

$$\bar{l}_{i\alpha} = \int_{\Omega_i} b\hat{\varphi}_\alpha \, dx = b_0 \int_{\Xi_i} \hat{\varphi}_\alpha(\xi) \omega(\xi) \, d\xi.$$

Therefore,

$$\bar{l}_i = b_0 h_1 h_2 \begin{pmatrix} \int_{\Xi_i} (1 - \xi_1)(1 - \xi_2) \, d\xi \\ \int_{\Xi_i} \xi_1(1 - \xi_2) \, d\xi \\ \int_{\Xi_i} \xi_1 \xi_2 \, d\xi \\ \int_{\Xi_i} (1 - \xi_1)\xi_2 \, d\xi \end{pmatrix} = \frac{b_0 h_1 h_2}{4} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

6.6 Let $\Xi = (0, 1)^2$ be a standard domain. With respect to $\alpha \in \{1, \dots, 4\}$, let $\hat{\varphi}_{(\alpha)}(\xi)$ are basis functions on Ξ . Here, the following holds:

$$\begin{aligned} \boldsymbol{\varepsilon}(\xi) &= \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ 2\varepsilon_{12} \end{pmatrix} = \begin{pmatrix} \frac{\partial u_{h1}}{\partial x_1} \\ \frac{\partial u_{h2}}{\partial x_1} \\ \frac{\partial u_{h2}}{\partial x_1} + \frac{\partial u_{h1}}{\partial x_2} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial \hat{\varphi}_1}{\partial x_1} & \frac{\partial \hat{\varphi}_2}{\partial x_1} & \frac{\partial \hat{\varphi}_3}{\partial x_1} & \frac{\partial \hat{\varphi}_4}{\partial x_1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{\partial \hat{\varphi}_1}{\partial x_2} & \frac{\partial \hat{\varphi}_2}{\partial x_2} & \frac{\partial \hat{\varphi}_3}{\partial x_2} & \frac{\partial \hat{\varphi}_4}{\partial x_2} \\ \frac{\partial \hat{\varphi}_1}{\partial x_2} & \frac{\partial \hat{\varphi}_2}{\partial x_2} & \frac{\partial \hat{\varphi}_3}{\partial x_2} & \frac{\partial \hat{\varphi}_4}{\partial x_2} & \frac{\partial \hat{\varphi}_1}{\partial x_1} & \frac{\partial \hat{\varphi}_2}{\partial x_1} & \frac{\partial \hat{\varphi}_3}{\partial x_1} & \frac{\partial \hat{\varphi}_4}{\partial x_1} \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{12} \\ u_{13} \\ u_{14} \\ u_{21} \\ u_{22} \\ u_{23} \\ u_{24} \end{pmatrix} \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\omega(\xi)} \begin{pmatrix} \frac{\partial \hat{x}_2}{\partial \xi_2} \frac{\partial \hat{\varphi}_1}{\partial \xi_1} - \frac{\partial \hat{x}_2}{\partial \xi_1} \frac{\partial \hat{\varphi}_1}{\partial \xi_2} & \frac{\partial \hat{x}_2}{\partial \xi_2} \frac{\partial \hat{\varphi}_2}{\partial \xi_1} - \frac{\partial \hat{x}_2}{\partial \xi_1} \frac{\partial \hat{\varphi}_2}{\partial \xi_2} \\ -\frac{\partial \hat{x}_1}{\partial \xi_2} \frac{\partial \hat{\varphi}_1}{\partial \xi_1} + \frac{\partial \hat{x}_1}{\partial \xi_1} \frac{\partial \hat{\varphi}_1}{\partial \xi_2} & 0 \\ \frac{\partial \hat{x}_2}{\partial \xi_2} \frac{\partial \hat{\varphi}_3}{\partial \xi_1} - \frac{\partial \hat{x}_2}{\partial \xi_1} \frac{\partial \hat{\varphi}_3}{\partial \xi_2} & \frac{\partial \hat{x}_2}{\partial \xi_2} \frac{\partial \hat{\varphi}_4}{\partial \xi_1} - \frac{\partial \hat{x}_2}{\partial \xi_1} \frac{\partial \hat{\varphi}_4}{\partial \xi_2} \\ 0 & 0 \\ -\frac{\partial \hat{x}_1}{\partial \xi_2} \frac{\partial \hat{\varphi}_3}{\partial \xi_1} + \frac{\partial \hat{x}_1}{\partial \xi_1} \frac{\partial \hat{\varphi}_3}{\partial \xi_2} & -\frac{\partial \hat{x}_1}{\partial \xi_2} \frac{\partial \hat{\varphi}_4}{\partial \xi_1} + \frac{\partial \hat{x}_1}{\partial \xi_1} \frac{\partial \hat{\varphi}_4}{\partial \xi_2} \\ 0 & 0 \\ -\frac{\partial \hat{x}_1}{\partial \xi_2} \frac{\partial \hat{\varphi}_1}{\partial \xi_1} + \frac{\partial \hat{x}_1}{\partial \xi_1} \frac{\partial \hat{\varphi}_1}{\partial \xi_2} & -\frac{\partial \hat{x}_1}{\partial \xi_2} \frac{\partial \hat{\varphi}_2}{\partial \xi_1} + \frac{\partial \hat{x}_1}{\partial \xi_1} \frac{\partial \hat{\varphi}_2}{\partial \xi_2} \\ \frac{\partial \hat{x}_2}{\partial \xi_2} \frac{\partial \hat{\varphi}_1}{\partial \xi_1} - \frac{\partial \hat{x}_2}{\partial \xi_1} \frac{\partial \hat{\varphi}_1}{\partial \xi_2} & \frac{\partial \hat{x}_2}{\partial \xi_2} \frac{\partial \hat{\varphi}_2}{\partial \xi_1} - \frac{\partial \hat{x}_2}{\partial \xi_1} \frac{\partial \hat{\varphi}_2}{\partial \xi_2} \\ -\frac{\partial \hat{x}_1}{\partial \xi_2} \frac{\partial \hat{\varphi}_3}{\partial \xi_1} + \frac{\partial \hat{x}_1}{\partial \xi_1} \frac{\partial \hat{\varphi}_3}{\partial \xi_2} & 0 \\ \frac{\partial \hat{x}_2}{\partial \xi_2} \frac{\partial \hat{\varphi}_3}{\partial \xi_1} - \frac{\partial \hat{x}_2}{\partial \xi_1} \frac{\partial \hat{\varphi}_3}{\partial \xi_2} & \frac{\partial \hat{x}_2}{\partial \xi_2} \frac{\partial \hat{\varphi}_4}{\partial \xi_1} - \frac{\partial \hat{x}_2}{\partial \xi_1} \frac{\partial \hat{\varphi}_4}{\partial \xi_2} \\ 0 & 0 \\ -\frac{\partial \hat{x}_1}{\partial \xi_2} \frac{\partial \hat{\varphi}_3}{\partial \xi_1} + \frac{\partial \hat{x}_1}{\partial \xi_1} \frac{\partial \hat{\varphi}_3}{\partial \xi_2} & -\frac{\partial \hat{x}_1}{\partial \xi_2} \frac{\partial \hat{\varphi}_4}{\partial \xi_1} + \frac{\partial \hat{x}_1}{\partial \xi_1} \frac{\partial \hat{\varphi}_4}{\partial \xi_2} \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{12} \\ u_{13} \\ u_{14} \\ u_{21} \\ u_{22} \\ u_{23} \\ u_{24} \end{pmatrix} \\
&= \mathbf{B}(\xi) \bar{\mathbf{u}}_i,
\end{aligned}$$

where $\omega(\xi) = \det(\partial_\xi \mathbf{x}^\top)$. The element coefficient matrix becomes

$$\mathbf{K}_i = \int_{\Omega_i} \mathbf{B}^\top(\mathbf{x}) \mathbf{D} \mathbf{B}(\mathbf{x}) \, d\mathbf{x} = \int_{\Xi} \mathbf{B}^\top(\xi) \mathbf{D} \mathbf{B}(\xi) \omega(\xi) \, d\xi.$$

Here, the integral of the right-hand side can be obtained by the Gaussian quadrature.

Chapter 8

8.1 When the θ -type elastic problem (Problem 8.9.2) was made into a state determination problem, a self-adjoint relationship was obtained with respect to the mean compliance f_0 defined by Eq. (8.9.6). Similarly, when the θ -type

Poisson problem (Problem 8.2.3) is made into a state determination problem, if

$$f_0(u) = \int_D b(\theta) u \, dx + \int_{\Gamma_N} p_N u \, d\gamma - \int_{\Gamma_D} \phi^\alpha(\theta) u_D \partial_\nu u \, d\gamma$$

is taken to be an objective function, the self-adjoint relationship is obtained. Moreover, the θ -derivative of f_0 becomes

$$\tilde{f}'_0(\theta)[\vartheta] = \langle g_0, \vartheta \rangle = \int_D \left(2b_\theta u - \alpha \phi^{\alpha-1} \phi_\theta \nabla u \cdot \nabla u \right) \vartheta \, dx.$$

8.2 The θ -type expanded Poisson problem becomes as below.

Problem P.8.1 (θ -Type Expanded Poisson Problem) Let D be a $d \in \{2, 3\}$ -dimensional Lipschitz domain. With respect to $\theta \in \mathcal{D}$, $b \in C^1(\mathcal{D}; L^{2q_R}(D; \mathbb{R}))$, $c_\Omega \in L^\infty(D; \mathbb{R})$, $p_B \in L^{2q_R}(\partial D; \mathbb{R})$, $c_{\partial\Omega} \in L^\infty(\partial D; \mathbb{R})$ are assumed to be given, where let $q_R > d$. Here, obtain $u : D \rightarrow \mathbb{R}$ that satisfies

$$\begin{aligned} -\nabla \cdot (\phi^\alpha(\theta) \nabla u) + c_\Omega u &= b(\theta) \quad \text{in } D, \\ \phi^\alpha(\theta) \partial_\nu u + c_{\partial\Omega} u &= p_B \quad \text{on } \partial D. \end{aligned}$$

□

Let the Lagrange function with respect to Problem P.8.1 be

$$\begin{aligned} \mathcal{L}_S(\theta, u, v) &= \int_D (-\phi^\alpha(\theta) \nabla u \cdot \nabla v - c_\Omega u v + b(\theta) v) \, dx \\ &\quad + \int_{\partial\Omega} (-c_{\partial\Omega} u v + p_B v) \, d\gamma \end{aligned}$$

by applying Problem 5.1.4. As an analogy with the mean compliance with respect to the θ -type linear elastic problem, let an objective function be

$$f_0(u) = \int_D b(\theta) u \, dx + \int_{\partial D} p_B u \, d\gamma, \quad (\text{P.8.1})$$

and a constraint function with respect to the domain measure be Eq. (8.9.7). Here, the θ -type topology optimization problem becomes as follows.

Problem P.8.2 (θ -Type Topology Optimization Problem) Let \mathcal{D} be Eq. (8.1.4), and $\mathcal{S} = W^{1,2q_R}(D; \mathbb{R})$. Let f_0 and f_1 be Eqs. (P.8.1) and (8.9.7), respectively. In this case, obtain θ satisfying

$$\min_{(\theta, u) \in \mathcal{D} \times \mathcal{S}} \{ f_0(\theta, u) \mid f_1(\theta) \leq 0, \text{ Problem P.8.1} \}.$$

□

In order to obtain the θ -derivative of f_0 , let the Lagrange function with respect to f_0 be

$$\begin{aligned}\mathcal{L}_0(\theta, u, v_0) &= f_0(\theta, u) + \mathcal{L}_S(\theta, u, v_0) \\ &= \int_D \{-\phi^\alpha(\theta) \nabla u \cdot \nabla v_0 + b(\theta)(u + v_0)\} dx \\ &\quad + \int_{\partial\Omega} p_B(u + v_0) d\gamma.\end{aligned}$$

Let the Fréchet derivative of \mathcal{L}_0 with respect to an arbitrary variation $(\vartheta, \hat{u}, \hat{v}_0) \in X \times U \times U$ (where $U = H^1(D; \mathbb{R})$) of (θ, u, v_0) be

$$\begin{aligned}\mathcal{L}'_0(\theta, u, v_0) [\vartheta, \hat{u}, \hat{v}_0] &= \mathcal{L}_{0\theta}(\theta, u, v_0) [\vartheta] + \mathcal{L}_{0u}(\theta, u, v_0) [\hat{u}] \\ &\quad + \mathcal{L}_{0v_0}(\theta, u, v_0) [\hat{v}_0].\end{aligned}\tag{P.8.2}$$

The third term on the right-hand side of Eq. (P.8.2) becomes

$$\mathcal{L}_{0v_0}(\theta, u, v_0) [\hat{v}_0] = \mathcal{L}_{Sv_0}(\theta, u, v_0) [\hat{v}_0] = \mathcal{L}_S(\theta, u, \hat{v}_0).$$

Moreover, the second term on the right-hand side of Eq. (P.8.2) becomes

$$\mathcal{L}_{0u}(\theta, u, v_0) [\hat{u}] = \mathcal{L}_S(\theta, \hat{u}, v_0).$$

Here, the self-adjoint relationship:

$$u = v_0$$

holds. Furthermore, the first term on the right-hand side of Eq. (P.8.2) becomes

$$\mathcal{L}_{0\theta}(\theta, u, v_0) [\vartheta] = \int_D \{b_\theta \cdot (u + v_0) - \alpha \phi^{\alpha-1} \phi_\theta \nabla u \cdot \nabla v_0\} \vartheta dx.$$

Hence, we get

$$\begin{aligned}\tilde{f}'_0(\theta) [\vartheta] &= \mathcal{L}_{0\theta}(\theta, u, v_0) [\vartheta] = \langle g_0, \vartheta \rangle \\ &= \int_D (2b_\theta \cdot u - \alpha \phi^{\alpha-1} \phi_\theta \nabla u \cdot \nabla u) \vartheta dx.\end{aligned}$$

On the other hand, the θ -derivative of $f_1(\theta)$ becomes

$$f'_1(\theta) [\vartheta] = \langle g_1, \vartheta \rangle = \int_D \phi_\theta \vartheta dx.$$

Here, the KKT conditions with respect to Problem P.8.2 are given as the conditions for which

$$\begin{aligned}\langle g_0 + \lambda_1 g_1, \vartheta \rangle &= \left\langle 2b_\theta \cdot u + \left(-\alpha \phi^{\alpha-1} \nabla u \cdot \nabla u + \lambda_1 \right) \phi_\theta, \vartheta \right\rangle = 0, \\ f_1(\theta) &\leq 0, \\ \lambda_1 f_1(\theta) &= 0, \\ \lambda_1 &\geq 0\end{aligned}$$

hold with respect to an arbitrary $\vartheta \in X$. Here, λ_1 is the Lagrange multiplier with respect to the domain measure constraint.

8.3 Let the Lagrange function with respect to Problem 8.12.1 be

$$\mathcal{L}(\theta, \beta, u, v_1, \dots, v_m, \lambda_1, \dots, \lambda_m) = \beta + \sum_{i \in \{1, \dots, m\}} \lambda_i \mathcal{L}_i(\theta, \beta, u, v_i),$$

where $\lambda = \{\lambda_1, \dots, \lambda_m\}^\top$ is a Lagrange multiplier with respect to $f_1 - \beta \leq 0, \dots, f_m - \beta \leq 0$, and

$$\mathcal{L}_i(\theta, \beta, u, v_i) = f_i(\theta, u) - \beta + \mathcal{L}_S(\theta, u, v_i).$$

Here, let \mathcal{L}_S be defined in Eq. (8.2.4). The Fréchet derivative of \mathcal{L} with respect to an arbitrary variation $(\vartheta, \hat{\beta}, \hat{u}, \hat{v}_1, \dots, \hat{v}_m) \in X \times \mathbb{R} \times U^{m+1}$ of $(\theta, \beta, u, v_1, \dots, v_m)$ is written as

$$\begin{aligned}\mathcal{L}'(\theta, \beta, u, v_1, \dots, v_m, \lambda_1, \dots, \lambda_m) &[\vartheta, \hat{\beta}, \hat{u}, \hat{v}_1, \dots, \hat{v}_m] \\ &= \mathcal{L}_\theta(\theta, \beta, u, v_1, \dots, v_m, \lambda_1, \dots, \lambda_m) [\vartheta] \\ &\quad + \mathcal{L}_\beta(\theta, \beta, u, v_1, \dots, v_m, \lambda_1, \dots, \lambda_m) [\hat{\beta}] \\ &\quad + \sum_{i \in \{1, \dots, m\}} \lambda_i \mathcal{L}_{iu}(\theta, \beta, u, v_i) [\hat{u}] \\ &\quad + \sum_{i \in \{1, \dots, m\}} \lambda_i \mathcal{L}_{iv_i}(\theta, \beta, u, v_i) [v'_i].\end{aligned}\tag{P.8.3}$$

The fourth term on the right-hand side of Eq. (P.8.3) becomes 0 when u is the weak solution of the state determination problem. The third term on the right-hand side of Eq. (P.8.3) becomes

$$\begin{aligned} & \sum_{i \in \{1, \dots, m\}} \lambda_i \mathcal{L}_{iu}(\theta, \beta, u, v_i) [\hat{u}] \\ &= \sum_{i \in \{1, \dots, m\}} \lambda_i (f_{iu}(\theta, u) [\hat{u}] + \mathcal{L}_{Su}(\theta, u, v_i) [\hat{u}]). \end{aligned}$$

When v_1, \dots, v_m are the weak solutions of adjoint problem (Problem 8.5.1) with respect to f_1, \dots and f_m , respectively, it becomes 0. The second term on the right-hand side of Eq. (P.8.3) becomes

$$\mathcal{L}_\beta(\theta, \beta, u, v_1, \dots, v_m, \lambda_1, \dots, \lambda_m) [\hat{\beta}] = (1 - \lambda_1 - \dots - \lambda_m) \hat{\beta}.$$

Furthermore, the first term on the right-hand side of Eq. (P.8.3) can be written as

$$\begin{aligned} & \mathcal{L}_\theta(\theta, \beta, u, v_1, \dots, v_m, \lambda_1, \dots, \lambda_m) [\vartheta] \\ &= \sum_{i \in \{1, \dots, m\}} \lambda_i \mathcal{L}_{i\theta}(\theta, \beta, u, v_i) [\vartheta] = \sum_{i \in \{1, \dots, m\}} \lambda_i \langle g_i, \vartheta \rangle. \end{aligned}$$

Here g_i is given by Eq. (8.5.6).

Hence, the KKT conditions with respect to Problem 8.12.1 are given as the conditions under which

$$\lambda_1 + \dots + \lambda_m = 1, \quad (\text{P.8.4})$$

$$\left\langle \sum_{i \in \{1, \dots, m\}} \lambda_i g_i, \vartheta \right\rangle = 0,$$

$$f_i(\theta) \leq 0 \quad \text{for } i \in \{1, \dots, m\},$$

$$\lambda_i f_i(\theta) = 0 \quad \text{for } i \in \{1, \dots, m\},$$

$$\lambda_i \geq 0 \quad \text{for } i \in \{1, \dots, m\}$$

holds with respect to an arbitrary $\vartheta \in X$.

Moreover, the solution to this problem using the gradient method with respect to constrained problems becomes as seen below. Imagine a situation with a simple algorithm (Algorithm 3.6) shown in Sect. 3.7.1, and suppose the replacements such as those shown in Sect. 8.7 are conducted. In this problem, g_0 (g_0 in Problem 3.7.1) becomes 0. Therefore $\vartheta_{g_0} = 0$. Moreover, set $\beta = \max_{i \in \{1, \dots, m\}} f_i - \epsilon$ with ϵ as a positive constant. Here, Eq. (8.7.3) for

obtaining the Lagrange multiplier becomes

$$(\langle g_i, \vartheta_{gj} \rangle)_{(i,j) \in I_A^2} (\lambda_j)_{j \in I_A} = -(f_i)_{i \in I_A}. \quad (\text{P.8.5})$$

If $(g_i)_{i \in I_A}$ is linearly independent, $(\lambda_j)_{j \in I_A}$ satisfying Eq. (P.8.5) is uniquely determined. Here, if $c = \sum_{j \in I_A} \lambda_j$ is used to replace $(\lambda_j/c)_{j \in I_A}$ with $(\lambda_j)_{j \in I_A}$ and $(c\vartheta_{gj})_{j \in I_A}$ with $(\vartheta_{gj})_{j \in I_A}$, Eqs. (P.8.4) and (P.8.5) are simultaneously satisfied. However, if Eq. (8.7.2) is used to seek ϑ_g , these replacements become unnecessary.

- 8.4** If \mathbf{u} is the solution of the state determination problem (Problem 8.9.2), it satisfies $\min_{\mathbf{u} \in U} \pi$ (Theorem 5.2.9). On the other hand, the maximum point with respect to θ of $\pi(\theta, \mathbf{u})$ becomes the minimum point of $-\pi(\theta, \mathbf{u})$. When \mathbf{u} is a solution of the state determination problem,

$$-\pi_\theta(\theta, \mathbf{u})[\vartheta] = \frac{1}{2} \langle g_0, \vartheta \rangle$$

holds with respect to an arbitrary $\vartheta \in X$. Here, g_0 represents a vector of Eq. (8.9.14).

- 8.5** If (\mathbf{u}, p) is the solution of a state determination problem (Problem 8.10.2), it satisfies $\min_{\mathbf{u} \in U} \max_{p \in P} \pi$ (Theorem 5.6.6). On the other hand, when (\mathbf{u}, p) is the solution of the state determination problem,

$$\pi_\theta(\theta, \mathbf{u}, p)[\vartheta] = \frac{1}{2} \langle g_0, \vartheta \rangle$$

holds with respect to an arbitrary $\vartheta \in X$. Here, g_0 represents a vector of Eq. (8.10.17).

Chapter 9

- 9.1** With respect to the second term on the right-hand side of Eq. (9.8.9),

$$\begin{aligned} & \left\| \left(\sum_{j \in \{1, \dots, d-1\}} \{ \boldsymbol{\tau}_j \cdot \nabla (p_N v_i) \} \boldsymbol{\tau}_j \right) \cdot \boldsymbol{\varphi} \right\|_{L^1(\Gamma_p(\boldsymbol{\phi}); \mathbb{R})} \\ & \leq (d-1) \max_{j \in \{1, \dots, d-1\}} \left(\|\boldsymbol{\tau}_j\|_{L^\infty(\Gamma_p(\boldsymbol{\phi}); \mathbb{R})}^2 \right. \\ & \quad \left. \times \|\nabla (p_N v_i)\|_{L^2(\Gamma_p(\boldsymbol{\phi}); \mathbb{R})} \right) \|\boldsymbol{\varphi}\|_{L^2(\Gamma_p(\boldsymbol{\phi}); \mathbb{R}^d)} \end{aligned} \quad (\text{P.9.1})$$

holds. Here,

$$\begin{aligned}
 \|\nabla(p_N v_i)\|_{L^2(\Gamma_p(\phi); \mathbb{R})} &\leq \|p_N v_i\|_{H^1(\Gamma_p(\phi); \mathbb{R})} \\
 &\leq \|p_N\|_{W^{1,4}(\Gamma_p(\phi); \mathbb{R})} \|v_i\|_{W^{1,4}(\Gamma_p(\phi); \mathbb{R})} \\
 &\leq \|\gamma_{\partial\Omega}\|^2 \|p_N\|_{C^{1,1}(D; \mathbb{R})} \|v_i\|_{W^{2,4}(D; \mathbb{R})}
 \end{aligned}$$

holds. Hence,

$$\begin{aligned}
 &(\text{Eq. (P.9.1) の右辺}) \\
 &\leq \|\gamma_{\partial\Omega}\|^3 (d-1) \max_{j \in \{1, \dots, d-1\}} \|\tau_j\|_{H^{3/2} \cap C^{0,1}(\Gamma_p(\phi); \mathbb{R})}^2 \\
 &\quad \times \|p_N\|_{C^{1,1}(D; \mathbb{R})} \|v_i\|_{W^{2,4}(D; \mathbb{R})} \|\varphi\|_X
 \end{aligned}$$

holds. If Hypothesis 9.5.1 is satisfied, the right-hand side of the equation above becomes bounded, and the second term on the right-hand side of Eq. (9.8.9) becomes an element of X' . Furthermore, from the fact that $\nabla(p_N v_i) = v_i \nabla p_N + p_N \nabla v_i \in W^{1,4}(D; \mathbb{R})$ and $\tau_j \in H^{3/2} \cap C^{0,1}(\Gamma_p(\phi); \mathbb{R})$, the second term on the right-hand side of Eq. (9.8.9) is included in $H^{1/2} \cap L^\infty(\Gamma_p(\phi); \mathbb{R}^d)$.

9.2 Let the Lagrange function of Problem 9.15.1 be

$$\mathcal{L}_S(\phi, u, v) = - \int_{\Omega(\phi)} \nabla u \cdot \nabla v \, dx + \int_{\partial\Omega(\phi)} (p_R v - c_{\partial\Omega} u v) \, d\gamma.$$

Moreover, the Lagrange function with respect to f_i is set to be

$$\begin{aligned}
 \mathcal{L}_i(\phi, u, v_i) &= f_i(\phi, u) + \mathcal{L}_S(\phi, u, v_i) \\
 &= - \int_{\Omega(\phi)} \nabla u \cdot \nabla v_i \, dx \\
 &\quad + \int_{\partial\Omega(\phi)} (\eta_{Ri}(\phi, u) + p_R v_i - c_{\partial\Omega} u v_i) \, d\gamma.
 \end{aligned}$$

Applying the formulae using the shape derivative of a function, the shape derivative of \mathcal{L}_i can be written as

$$\begin{aligned}
 \mathcal{L}'_i(\phi, u, v_i) [\varphi, \hat{u}, \hat{v}_i] &= \mathcal{L}_{i\phi'}(\phi, u, v_i) [\varphi] + \mathcal{L}_{iu}(\phi, u, v_i) [\hat{u}] + \mathcal{L}_{iv_i}(\phi, u, v_i) [\hat{v}_i].
 \end{aligned} \tag{P.9.2}$$

The third term on the right-hand side of Eq. (P9.2) becomes

$$\mathcal{L}_{i v_i}(\phi, u, v_i)[\hat{v}_i] = \mathcal{L}_{S v_i}(\phi, u, v_i)[\hat{v}_i] = \mathcal{L}_S(\phi, u, \hat{v}_i).$$

If u is a weak solution of the state determination problem (Problem 9.15.1), it becomes 0. Moreover, the second term on the right-hand side of Eq. (P9.2) becomes

$$\begin{aligned} & \mathcal{L}_{i u}(\phi, u, v_i)[\hat{u}] \\ &= - \int_{\Omega(\phi)} \nabla \hat{u} \cdot \nabla v_i \, dx + \int_{\partial\Omega(\phi)} (\eta_{R i u}(\phi, u)[\hat{u}] - c_{\partial\Omega} v_i \hat{u}) \, d\gamma. \end{aligned}$$

When v_i is a weak solution of an adjoint problem with respect to f_i such as the following, the second term on the right-hand side of Eq. (P9.2) becomes 0 too.

Problem P9.1 (Adjoint Problem with Respect to f_i) When a solution u of Problem 9.15.1 with respect to $\phi \in \mathcal{D}$ is given, obtain $v_i : \Omega(\phi) \rightarrow \mathbb{R}$ which satisfies

$$\begin{aligned} -\Delta v_i &= 0 \quad \text{in } \Omega(\phi), \\ \partial_\nu v_i + c_{\partial\Omega}(\phi) v_i &= \eta_{R i u}(\phi, u) \quad \text{on } \partial\Omega(\phi). \end{aligned} \quad \square$$

Furthermore, the first term on the right-hand side of Eq. (P9.2) becomes

$$\begin{aligned} & \mathcal{L}_{i \phi'}(\phi, u, v_i)[\varphi] \\ &= \int_{\Omega(\phi)} \left\{ \nabla u \cdot (\nabla \varphi^\top \nabla v_i) + \nabla v_i \cdot (\nabla \varphi^\top \nabla u) \right. \\ & \quad \left. - (\nabla u \cdot \nabla v_i) \nabla \cdot \varphi \right\} dx \\ &+ \int_{\partial\Omega(\phi)} \left\{ \kappa (\eta_{R i}(\phi, u) + p_R v_i - c_{\partial\Omega} u v_i) \mathbf{v} \cdot \varphi \right. \\ & \quad \left. - \nabla_\tau (\eta_{R i}(\phi, u) + p_R v_i - c_{\partial\Omega} u v_i) \cdot \varphi_\tau \right\} d\gamma \\ &+ \int_{\Theta(\phi)} (\eta_{R i}(\phi, u) + p_R v_i - c_{\partial\Omega} u v_i) \boldsymbol{\tau} \cdot \varphi \, d\zeta. \end{aligned}$$

In order to obtain this integral, the fact that $\partial\Omega(\phi)$ is piecewise $H^3 \cap C^{1,1}$ was used. Moreover, the known function was assumed to be fixed with the material.

With the above results in mind, if u and v_i are assumed to be the weak solutions of Problems 9.15.1 and P9.1,

$$\begin{aligned}\tilde{f}'_i(\boldsymbol{\phi})[\boldsymbol{\varphi}] &= \mathcal{L}_{i\boldsymbol{\phi}'}(\boldsymbol{\phi}, u, v_i)[\boldsymbol{\varphi}] = \langle \mathbf{g}_i, \boldsymbol{\varphi} \rangle \\ &= \int_{\Omega(\boldsymbol{\phi})} \left(\mathbf{G}_{\Omega i} \cdot \nabla \boldsymbol{\varphi}^\top + g_{\Omega i} \nabla \cdot \boldsymbol{\varphi} \right) dx + \int_{\partial\Omega(\boldsymbol{\phi})} \mathbf{g}_{\partial\Omega i} \cdot \boldsymbol{\varphi} dy \\ &\quad + \int_{\Theta(\boldsymbol{\phi})} \mathbf{g}_{\Theta i} \cdot \boldsymbol{\varphi} dy\end{aligned}$$

can be written. Here, we get

$$\begin{aligned}\mathbf{G}_{\Omega i} &= \nabla u (\nabla v_i)^\top + \nabla v_i (\nabla u)^\top, \\ g_{\Omega i} &= -\nabla u \cdot \nabla v_i, \\ \mathbf{g}_{\partial\Omega i} &= \kappa (\eta_{Ri}(\boldsymbol{\phi}, u) + p_R v_i - c_{\partial\Omega} u v_i) \boldsymbol{\tau} \\ &\quad - \sum_{j \in \{1, \dots, d-1\}} \{ \boldsymbol{\tau}_j \cdot \nabla (\eta_{Ri}(\boldsymbol{\phi}, u) + p_R v_i - c_{\partial\Omega} u v_i) \} \boldsymbol{\tau}_j, \\ \mathbf{g}_{\Theta i} &= (\eta_{Ri}(\boldsymbol{\phi}, u) + p_R v_i - c_{\partial\Omega} u v_i) \boldsymbol{\tau}.\end{aligned}$$

The similar regularity for \mathbf{g}_i in Theorem 9.8.2 means $\mathbf{G}_{\Omega i} \in H^1 \cap L^\infty(\Omega(\boldsymbol{\phi}); \mathbb{R}^{d \times d})$, $g_{\Omega i} \in H^1 \cap L^\infty(\Omega(\boldsymbol{\phi}); \mathbb{R})$ and $\mathbf{g}_{\partial\Omega i} \in H^{1/2} \cap L^\infty(\partial\Omega(\boldsymbol{\phi}); \mathbb{R}^d)$. To obtain the results, from the proof of Theorem 9.8.2, considering that u and v_i are elements of $W^{2,4}(D; \mathbb{R})$, the regularity of known function required in this case is

$$\begin{aligned}c_{\partial\Omega} C_{S'}^1(B; C^{1,1}(D; \mathbb{R})), \quad p_R \in C_{S'}^1(B; C^{1,1}(D; \mathbb{R})), \\ \eta_{Ri}(\boldsymbol{\phi}, u) \in W^{2,q_R}(D; \mathbb{R}), \quad \eta_{Riu}(\boldsymbol{\phi}, u)[\hat{u}] \in W^{1,4}(D; \mathbb{R})\end{aligned}$$

in a neighborhood $B \subset Y$ of $\boldsymbol{\phi} \in \mathcal{D}^\circ$. On the other side, with respect to an opening angle β of a corner point, the condition $\beta < 2\pi/3$ when the corner point is between boundaries of the same type will be applied.

- 9.3 Let us use Eq. (9.15.3) in order to obtain \hat{g}_{iC} . With respect to the first term in the right-hand integrand of Eq. (9.15.3),

$$\begin{aligned}\nabla u &= \left(\cos \theta \frac{\partial}{\partial r} - \frac{\sin \theta}{r} \frac{\partial}{\partial \theta} \right) u = \frac{k_j}{2\epsilon^{1/2}} \begin{pmatrix} \cos(\theta/2) \\ \sin(\theta/2) \end{pmatrix}, \\ \nabla v_i &= \frac{l_{ij}}{2\epsilon^{1/2}} \begin{pmatrix} \cos(\theta/2) \\ \sin(\theta/2) \end{pmatrix}\end{aligned}$$

holds. Here, we obtain

$$\nabla u \cdot \nabla v_i = \frac{k_j l_{ij}}{4\epsilon}.$$

Substituting this result into the first term of the right-hand integrand of Eq. (9.15.3) gives

$$\begin{aligned} & - \int_0^{2\pi} (\nabla u \cdot \nabla v_i) \mathbf{v} \cdot \boldsymbol{\varphi} \epsilon \, d\theta \\ &= \int_0^{2\pi} \frac{k_j l_{ij}}{4} (\varphi_1 \cos \theta + \varphi_2 \sin \theta) \, d\theta = 0 \end{aligned} \quad (\text{P.9.3})$$

with respect to an arbitrary $\boldsymbol{\varphi} = (\varphi_1, \varphi_2)^\top \in \mathbb{R}^2$. Furthermore, with respect to the second term of the integrand,

$$\begin{aligned} \partial_v u &= \mathbf{v} \cdot \nabla u = \frac{k_j}{2\epsilon^{1/2}} \begin{pmatrix} -\cos \theta \\ -\sin \theta \end{pmatrix} \cdot \begin{pmatrix} \cos(\theta/2) \\ \sin(\theta/2) \end{pmatrix} \\ &= -\frac{k_j}{2\epsilon^{1/2}} \cos(\theta/2), \\ \partial_v u \nabla v_i &= -\frac{k_j l_{ij}}{4\epsilon} \cos(\theta/2) \begin{pmatrix} \cos(\theta/2) \\ \sin(\theta/2) \end{pmatrix} \end{aligned}$$

is established. Here the second term of the integrand becomes

$$\begin{aligned} \int_0^{2\pi} \partial_v u \nabla v_i \cdot \boldsymbol{\varphi} \epsilon \, d\theta &= \int_0^{2\pi} \partial_v v_i \nabla u \cdot \boldsymbol{\varphi} \epsilon \, d\theta \\ &= -\frac{k_j l_{ij}}{4} \begin{pmatrix} \pi \\ 0 \end{pmatrix} \cdot \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix} \end{aligned} \quad (\text{P.9.4})$$

with respect to an arbitrary $\boldsymbol{\varphi} = (\varphi_1, \varphi_2)^\top \in \mathbb{R}^2$. The same result holds for the third term of the integrand. Hence, from Eqs. (P.9.3) and (P.9.4),

$$\langle \hat{\mathbf{g}}_{iC}, \boldsymbol{\varphi} \rangle = -\frac{k_j l_{ij}}{2} \begin{pmatrix} \pi \\ 0 \end{pmatrix} \cdot \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix} \quad (\text{P.9.5})$$

can be obtained. From Eq. (P.9.5), we see that the shape derivative $\hat{\mathbf{g}}_{iC}$ with respect to a variation of a crack point is in the direction of the crack surface.

$\hat{\mathbf{g}}_{iM}$ becomes as follows. With respect to the first term on the right-hand integrand of Eq. (9.15.3),

$$\nabla u = \begin{pmatrix} \cos \theta \frac{\partial}{\partial r} - \frac{\sin \theta}{r} \frac{\partial}{\partial \theta} \\ \sin \theta \frac{\partial}{\partial r} + \frac{\cos \theta}{r} \frac{\partial}{\partial \theta} \end{pmatrix} u = \frac{k_j}{2\epsilon^{1/2}} \begin{pmatrix} -\sin(\theta/2) \\ \cos(\theta/2) \end{pmatrix},$$

$$\nabla v_i = \frac{l_{ij}}{2\epsilon^{1/2}} \begin{pmatrix} -\sin(\theta/2) \\ \cos(\theta/2) \end{pmatrix}$$

holds. Hence,

$$\nabla u \cdot \nabla v_i = \frac{k_j l_{ij}}{4\epsilon}$$

is obtained. If this result is substituted into the first term of the right-hand integrand of Eq. (9.15.3), it becomes

$$-\int_0^\pi (\nabla u \cdot \nabla v_i) \mathbf{v} \cdot \boldsymbol{\varphi} \epsilon d\theta = \int_0^\pi \frac{k_j l_{ij}}{4} (\varphi_1 \cos \theta + \varphi_2 \sin \theta) d\theta$$

$$= \frac{k_j l_{ij}}{2} \varphi_2 \quad (\text{P.9.6})$$

with respect to an arbitrary $\boldsymbol{\varphi} = (\varphi_1, \varphi_2)^\top \in \mathbb{R}^2$. Furthermore,

$$\partial_v u = \mathbf{v} \cdot \nabla u = \frac{k_j}{2\epsilon^{1/2}} \begin{pmatrix} -\cos \theta \\ -\sin \theta \end{pmatrix} \cdot \begin{pmatrix} -\sin(\theta/2) \\ \cos(\theta/2) \end{pmatrix}$$

$$= -\frac{k_j}{2\epsilon^{1/2}} \sin(\theta/2),$$

$$\partial_v u \nabla v_i = -\frac{k_j l_{ij}}{4\epsilon} \sin(\theta/2) \begin{pmatrix} -\sin(\theta/2) \\ \cos(\theta/2) \end{pmatrix}$$

holds. Here, the second term of the integrand becomes

$$\int_0^\pi \partial_v u \nabla v_i \cdot \boldsymbol{\varphi} \epsilon d\theta = \int_0^\pi \partial_v v_i \nabla u \cdot \boldsymbol{\varphi} \epsilon d\theta$$

$$= \frac{k_j l_{ij}}{8} \begin{pmatrix} \pi \\ -2 \end{pmatrix} \cdot \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix} \quad (\text{P.9.7})$$

with respect to an arbitrary $\boldsymbol{\varphi} = (\varphi_1, \varphi_2)^\top \in \mathbb{R}^2$. The third term of the integrand gives the same result. Hence, from Eqs. (P.9.6) and (P.9.7),

$$\langle \hat{\mathbf{g}}_{iM}, \boldsymbol{\varphi} \rangle = \frac{k_j l_{ij}}{4} \begin{pmatrix} \pi \\ 0 \end{pmatrix} \cdot \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix} \quad (\text{P.9.8})$$

can be obtained. Equation (P.9.8) shows that the shape derivative $\hat{\mathbf{g}}_{iM}$ at a point of a mixed boundary on a smooth boundary is in the direction of the Neumann boundary.

Afterword

In this book, the shape optimization problems of elastic bodies or flow fields are taken to be function optimizing problems, and explanations regarding their formulations and solutions are provided. Before reaching the goals, the following matters were discussed based on the author's knowledge:

- (1) What sorts of structures the optimal design problems have.
- (2) What the optimization theories guarantee.
- (3) What it means that the function space on which a function optimization problem is defined is a vector space.
- (4) What sort of structures the boundary value problems of elliptic partial differential equations have, and how numerical solutions of such problems can be obtained, as well as what sort of theorems guarantee their regularity.

In order to obtain the results in this book, support was received from many people. The idea for the traction method which became the basis of the H^1 gradient method is one that I thought of when I was visiting at Michigan University for ten months from November of 1991 as a visiting scholar of the Ministry of Education of Japan. During this time, the facts that Professor Noboru Kikuchi introduced me to important literature, and that (the late) Professor John E. Taylor approved my participation in his class "Structural Optimization" gave me the opportunity to develop the traction method. I first theorized the traction method to compensate a flaw I discovered in the growth strain method that I had researched prior to my research visit abroad. After returning home, when going through a solitary struggle to explain its idea as a gradient method in a function space, I was introduced to Professor Satoshi Kaizu by Professor Fumio Kikuchi and Professor Ichiro Hagiwara. Since then, Professor Kaizu has given me various instructions regarding function analysis with respect to this problem. Professor Fumio Kikuchi taught me the role of function analysis using real examples via the mathematics of the finite element method. Professor Masahisa Tabata gave me instructions whenever the opportunity arose concerning the importance of defining the Fréchet derivative as

an element of dual space. This forms the framework with respect to the derivatives in this book. Furthermore, I want to thank Professor Kohji Ohtsuka and Professor Masato Kimura for clarifying the link between a generalized J integral and a shape derivative. By their suggestion, I was able to obtain the shape derivatives of domain integral type. Professor Daisuke Tagami gave me the opportunity to take an intensive course, which virtually follows the contents of this book, and suggestions which helped me in many ways. Moreover, the error estimations in Chaps. 8 and 9 are based on Doctor Daisuke Murai's concept. Doctor Yusuke Naritomi helped me with the theory and numerical verification regarding the second-order derivatives. In regards to the existence issues of the optimum solutions in Chaps. 7 to 9, the discussion with Doctor Julius Fergy T. Rabago and Doctor Masayuki Aino formed the bases of the theories. I would like to express my gratitude here for their contribution.

Furthermore, Doctor Kenzen Takeuchi's contribution was a decisive factor in the development of actual programs. At present, Doctor Takeuchi is the lead developer for a program of structural optimum design software OPTISHAPE TS (Quint Corporation). The verifications of the theories and methods shown in this book were done using this program. Figure 2 in the Preface, Figs. 8.5 and 9.1 showing the results from the H^1 gradient method and numerical instability phenomenon were provided by Doctor Takeuchi. Moreover, Fig. 1 in the Preface was provided by Ms. Mai Nonogawa. Furthermore, the programs used in the numerical examples in Chaps. 8 and 9 were created by students of Azegami laboratories. I would like to express our gratitude here for their contribution.

In this book, a focus has been given to a method of formulating and finding solutions to shape optimization problems. However, the topology optimization problems and shape optimization problems discussed in this book are applied in a variety of optimum design problems in engineering. In engineering, a lot of innovation is required in order to determine the appropriate state determination problem and cost functions with respect to various topics, and to seek the equations to calculate the derivative with respect to each cost function theoretically. I would like to summarize these concepts on another occasion.

Nagoya, Japan

June 2020

Hideyuki Azegami

References

1. N. Aage, T.H. Poulsen, A. Gersborg-Hansen, O. Sigmund, Topology optimization of large scale Stokes flow problems. *Struct. Multidiscip. Optim.* **35**, 175–180 (2008)
2. R.A. Adams, J.J.F. Fournier, *Sobolev Spaces*, 2nd edn. (Academic Press, Amsterdam; Tokyo, 2003)
3. G. Allaire, *Shape Optimization by the Homogenization Method* (Springer, New York, 2002)
4. G. Allaire, F. Jouve, A.M. Toader, Structural optimization using sensitivity analysis and a level-set method. *J. Comput. Phys.* **194**, 363–393 (2004)
5. L. Armijo, Minimization of functions having Lipschitz-continuous first partial derivatives. *Pac. J. Math.* **16**, 1–3 (1966)
6. J.S. Arora, *Introduction to Optimum Design* (McGraw-Hill, New York; Tokyo, 1989)
7. M. Avellaneda, Optimal bounds and microgeometries for elastic two-phase composites. *SIAM J. Appl. Math.* **47**, 1216–1228 (1987)
8. H. Azegami, A solution to domain optimization problems (in Japanese). *Trans. Jpn. Soc. Mech. Eng. A* **60**(574), 1479–1486 (1994)
9. H. Azegami, Regularized solution to shape optimization problem (in Japanese). *Trans. Jpn. Soc. Industrial Appl. Math.* **23**(2), 83–138 (2014)
10. H. Azegami, S. Fukumoto, T. Aoyama, Shape optimization of continua using NURBS as basis functions. *Struct. Multidiscip. Optim.* **47**(2), 247–258 (2013)
11. H. Azegami, S. Kaizu, K. Takeuchi, Regular solution to topology optimization problems of continua. *JSIAM Lett.* **3**, 1–4 (2011)
12. H. Azegami, K. Ohtsuka, M. Kimura, Shape derivative of cost function for singular point: Evaluation by the generalized J integral. *JSIAM Lett.* **6**, 29–32 (2014)
13. H. Azegami, Y. Sugai, M. Shimoda, Shape optimization with respect to buckling (in Japanese). *Trans. Jpn. Soc. Mech. Eng. A* **66**(647), 1262–1267 (2000)
14. H. Azegami, K. Takeuchi, A smoothing method for shape optimization: Traction method using the Robin condition. *Int. J. Comput. Methods* **3**(1), 21–33 (2006)
15. H. Azegami, Z.C. Wu, Domain optimization analysis in linear elastic problems: Approach using traction method (in Japanese). *Trans. Jpn. Soc. Mech. Eng. A* **60**(578), 2312–2318 (1994)
16. H. Azegami, S. Yokoyama, E. Katamine, Solution to shape optimization problems of continua on thermal elastic deformation, in *Inverse Problems in Engineering Mechanics III*, ed. by M. Tanaka, G.S. Dulikravich (Elsevier, Tokyo, 2002), pp. 61–66
17. H. Azegami, L. Zhou, K. Umemura, N. Kondo, Shape optimization for a link mechanism. *Struct. Multidiscip. Optim.* **48**(1), 115–125 (2013)

18. N.V. Banichuk, Optimality conditions and analytical methods of shape optimization, in *Optimization of Distributed Parameter Structures*, vol. 2, ed. by E.J. Haug, J. Cea (Sijthoff & Noordhoff, Alphen aan den Rijn, 1981), pp. 973–1004
19. N.V. Banichuk, *Problems and Methods of Optimal Structural Design* (Plenum Press, New York, 1983)
20. N.V. Banichuk, *Introduction to Optimization of Structures* (Springer, New York, 1990)
21. R.G. Bartle, *The Elements of Real Analysis*, 2nd edn. (Wiley, New York, 1976)
22. M.S. Bazaraa, C.M. Shetty, *Nonlinear Programming: Theory and Algorithms* (Wiley, New York, 1979)
23. A.D. Belegundu, S.D. Rajan, A shape optimization approach based on natural design variables and shape functions. *Comput. Methods Appl. Mech. Eng.* **66**, 87–106 (1988)
24. M.P. Bendsøe, Optimal shape design as a material distribution problem. *Structural Optimization* **1**, 193–202 (1989)
25. M.P. Bendsøe, N. Kikuchi, Generating optimal topologies in structural design using a homogenization method. *Comput. Methods Appl. Mech. Eng.* **71**, 197–224 (1988)
26. M.P. Bendsøe, O. Sigmund, *Topology Optimization: Theory, Methods and Applications* (Springer, Berlin; Tokyo, 2003)
27. M.S. Berger, *Nonlinearity and Functional Analysis: Lectures on Nonlinear Problems in Mathematical Analysis* (Academic Press, New York, 1977)
28. T. Borrivall, J. Petersson, Topology optimization of fluids in Stokes flow. *Int. J. Numer. Methods Fluids* **41**, 77–107 (2003)
29. A. Boulkhemair, A. Chakib, A. Nachaoui, Uniform trace theorem and application to shape optimization. *Appl. Comput. Math.* **7**, 192–205 (2008)
30. V. Braibant, C. Fleury, Shape optimal design using B-splines. *Comput. Methods Appl. Mech. Eng.* **44**, 247–267 (1984)
31. V. Braibant, C. Fleury, An approximation concepts approach to shape optimal design. *Comput. Methods Appl. Mech. Eng.* **53**, 119–148 (1985)
32. H. Brezis, *Analyse Fonctionnelle: Théorie et Applications* (Masson, Paris, 1983)
33. F. Brezzi, M. Fortin, *Mixed and Hybrid Finite Element Methods* (Springer, New York; Tokyo, 1991)
34. J. Cea, Numerical methods of shape optimal design, in *Optimization of Distributed Parameter Structures*, vol. 2, ed. by E.J. Haug, J. Cea (Sijthoff & Noordhoff, Alphen aan den Rijn, The Netherlands, 1981), pp. 1049–1088
35. J. Cea, Problems of shape optimization, in *Optimization of Distributed Parameter Structures*, vol. 2, ed. by E.J. Haug, J. Cea (Sijthoff & Noordhoff, Alphen aan den Rijn, The Netherlands, 1981), pp. 1005–1048
36. D. Chénais, On the existence of a solution in a domain identification problem. *J. Math. Anal. Appl.* **52**, 189–219 (1975)
37. K.K. Choi, Shape design sensitivity analysis of displacement and stress constraints. *J. Struct. Mech.* **13**, 27–41 (1985)
38. K.K. Choi, E.J. Haug, Shape design sensitivity analysis of elastic structures. *J. Struct. Mech.* **11**, 231–269 (1983)
39. K.K. Choi, N.H. Kim, *Structural Sensitivity Analysis and Optimization, 1 & 2* (Springer, New York, 2005)
40. E.K.P. Chong, S.H. Źak, *An Introduction to Optimization* (Wiley, New York, 2008)
41. P.G. Ciarlet, *Three-Dimensional Elasticity* (North-Holland, Amsterdam; Tokyo, 1988)
42. P.G. Ciarlet, *Finite Element Methods*, in *Handbook of Numerical Analysis*, ed. by P.G. Ciarlet, J.L. Lions (Elsevier, Amsterdam; Tokyo: North-Holl, 1991)
43. M.C. Delfour, J.P. Zolésio, Tangent calculus and shape derivatives, in *Shape Optimization and Optimal Design: Proceedings of the IFIP Conference*, ed. by J. Cagnol, M.P. Polis, J.P. Zolésio (Marcel Dekker, New York; Basel, 2001), pp. 37–60
44. M.C. Delfour, J.P. Zolésio, *Shapes and Geometries: Metrics, Analysis, Differential Calculus, and Optimization*, 2nd edn. (Society for Industrial and Applied Mathematics, Philadelphia, 2011)

45. A.R. Diaz, N. Kikuchi, Solutions to shape and topology eigenvalue optimization problems using a homogenization method. *Int. J. Numer. Methods Eng.* **35**, 1487–1502 (1992)
46. A.R. Diaz, O. Sigmund, Checkerboard patterns in layout optimization. *Structural Optimization* **10**, 40–45 (1995)
47. L.C. Evans, *Partial Differential Equations*, 4th edn. (American Mathematical Society, Providence, RI, 2002)
48. L.C. Evans, R.F. Gariepy, *Measure Theory and Fine Properties of Functions* (CRC Press, Boca Raton, 1992)
49. A. Evgrafov, Topology optimization of slightly compressible fluids. *J. Appl. Math. Mech./Zeitschrift für Angewandte Mathematik und Mechanik* **86**, 46–62 (2006)
50. H. Fujita, H. Konno, K. Tanabe, *Optimization Method (in Japanese)* (Iwanami Shoten, Tokyo, 1994)
51. M. Fukushima, *Basics of Nonlinear Optimization (in Japanese)* (Asakura Publishing, Tokyo, 2001)
52. I.M. Gelfand, S.V. Fomin, R.A. Rev., English ed. by Silverman, *Calculus of Variations* (Prentice-Hall, Englewood Cliffs, NJ, 1963)
53. A. Gersborg-Hansen, O. Sigmund, R.B. Haber, Topology optimization of channel flow problems. *Struct. Multidiscip. Optim.* **3**, 181–192 (2005)
54. G.T. Gilbert, Positive definite matrices and Sylvester's criterion. *Am. Math. Monthly* **98**(1), 44–46 (1991)
55. V. Girault, P.A. Raviart, *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms* (Springer, Berlin; Tokyo, 1986)
56. P. Grisvard, *Elliptic Problems in Nonsmooth Domains* (Pitman Advanced Pub. Program, Boston, 1985)
57. J.K. Guest, J.H. Prévost, Topology optimization of creeping fluid flows using a Darcy-Stokes finite element. *Int. J. Numer. Methods Eng.* **66**, 461–484 (2006)
58. J.K. Guest, J.H. Prévost, Design of maximum permeability material structures. *Comput. Methods Appl. Mech. Eng.* **196**, 1006–1017 (2007)
59. O. Güler, *Foundations of Optimization* (Springer, New York, 2010)
60. J. Hadamard, *Mémoire des Savants Étrangers. Oeuvres de J. Hadamard*, chapter Mémoire sur le problème d'analyse relatif à l'équilibre des plaques élastiques encastrées, Mémoire des savants étrangers, pp. 515–629 (CNRS, Paris, 1968)
61. R.T. Haftka, Z. Gurdal, *Elements of Structural Optimization. 3rd Rev. and Expanded Ed.* (Kluwer Academic Publishers, Dordrecht, 1992)
62. J. Haslinger, R.A.E. Mäkinen, *Introduction to Shape Optimization: Theory, Approximation, and Computation* (SIAM, Philadelphia, 2003)
63. J. Haslinger, P. Neittaanmäki, *Finite Element Approximation for Optimal Shape Design: Theory and Application* (Wiley, Chichester, 1988)
64. J. Haslinger, P. Neittaanmäki, *Finite Element Approximation for Optimal Shape, Material and Topology Design*, 2nd edn. (Wiley, Chichester, 1996)
65. E.J. Haug, K.K. Choi, V. Komkov, *Design Sensitivity Analysis of Structural Systems* (Academic Press, Orlando, 1986)
66. F. Hecht, New development in freefem++. *J. Numer. Math.* **20**(3-4), 251–265 (2012)
67. A. Henrot, M. Pierre, in *Shape Variation and Optimization: A Geometrical Analysis*. Tracts in Mathematics, vol. 28 (European Mathematical Society, Zürich, 2018)
68. V. Horák, *Inverse Variational Principles of Continuum Mechanics* (Academia, nakladatelství Československé akademie věd, Praha, 1969)
69. H. Ihara, H. Azegami, M. Shimoda, Shape optimization for displacement path control problem considering geometrical non-linearity (in Japanese). *Trans. Jpn. Soc. Mech. Eng. A* **67**(656), 611–617 (2001)
70. H. Ihara, H. Azegami, M. Shimoda, K. Watanabe, Solution to shape optimization problem considering material non-linearity (in Japanese). *Trans. Jpn. Soc. Mech. Eng. A* **66**(646), 1111–1118 (2000)

71. H. Ihara, M. Shimoda, H. Azegami, T. Sakurai, Shape design by integrating shape optimization with topology optimization for multiobjective structures: an approach using homogenization method and traction method (in Japanese). *Trans. Jpn. Soc. Mech. Eng. A* **62**(596), 1091–1097 (1996)
72. H. Ihara, M. Shimoda, H. Azegami, T. Sakurai, Numerical analysis method of controlling deformation mode using topology-shape optimization technique (in Japanese). *Trans. Soc. Automot. Eng. Jpn.* **29**(1), 117–122 (1998)
73. M.H. Imam, Three-dimensional shape optimization. *Int. J. Numer. Methods Eng.* **18**, 661–673 (1982)
74. A.R. Inzarulfaisham, H. Azegami, Shape optimization of linear elastic continua for moving nodes of natural vibration modes to assigned positions and its application to chassis-like frame structures. *Trans. Jpn. Soc. Comput. Eng. Sci.* **7**, 43–50 (2004)
75. A.R. Inzarulfaisham, H. Azegami, Solution to boundary shape optimization problem of linear elastic continua with prescribed natural vibration mode shapes. *Struct. Multidiscip. Optim.* **27**(3), 210–217 (2004)
76. T. Iwai, A. Sugimoto, T. Aoyama, H. Azegami, Shape optimization problem of elastic bodies for controlling contact pressure. *JSIAM Lett.* **2**, 1–4 (2010)
77. Y. Iwata, H. Azegami, T. Aoyama, E. Katamine, Numerical solution to shape optimization problems for non-stationary Navier-Stokes problems. *JSIAM Lett.* **2**, 37–40 (2010)
78. J. Jahn, *Introduction to the Theory of Nonlinear Optimization*. 2nd Rev. edn. (Springer, Berlin; New York, 1996)
79. M.W. Jeter, *Mathematical Programming: An Introduction to Optimization* (M. Dekker, New York, 1986)
80. S. Kaizu, H. Azegami, Optimal shape problems and traction method (in Japanese). *Trans. Jpn. Soc. Ind. Appl. Math.* **16**(3), 277–290 (2006)
81. E. Katamine, H. Azegami, Solution to viscous flow field domain optimization problems: Approach by the traction method (in Japanese). *Trans. Jpn. Soc. Mech. Eng. B* **60**(579), 3859–3866 (1994)
82. E. Katamine, H. Azegami, Domain optimization analysis of potential flow field (in Japanese). *Trans. Jpn. Soc. Mech. Eng. B* **61**(581), 103–108 (1995)
83. E. Katamine, H. Azegami, Domain optimization analysis of viscous flow field: In the case of considering convective term (in Japanese). *Trans. Jpn. Soc. Mech. Eng. B* **61**(585), 1646–1653 (1995)
84. E. Katamine, H. Azegami, M. Hirai, Solution of shape identification problems on thermoelastic solids. *Int. J. Comput. Methods* **3**(3), 279–293 (2006)
85. E. Katamine, H. Azegami, M. Kojima, Boundary shape determination on steady-state heat conduction fields (in Japanese). *Trans. Jpn. Soc. Mech. Eng. B* **65**(629), 275–281 (1999)
86. E. Katamine, H. Azegami, Y. Mataura, Solution to shape identification problem on unsteady heat-conduction fields (in Japanese). *Trans. Jpn. Soc. Mech. Eng. B* **66**(641), 227–234 (2000)
87. E. Katamine, H. Azegami, T. Tsubata, S. Itoh, Solution to shape optimization problems of viscous flow fields. *Int. J. Comput. Fluid Dyn.* **19**(1), 45–51 (2005)
88. E. Katamine, H. Azegami, S. Yamaguchi, Shape identification analyses of potential flow field: Prescribed problems of pressure distribution and solution by the traction method (in Japanese). *Trans. Jpn. Soc. Mech. Eng. B* **64**(620), 1063–1070 (1998)
89. E. Katamine, Y. Iwata, H. Azegami, Shape optimization of un-steady heat-conduction fields for thermal dissipation maximization (in Japanese). *Trans. Jpn. Soc. Mech. Eng. B* **74**(743), 1609–1616 (2008)
90. E. Katamine, Y. Kawase, H. Azegami, Shape optimization of forced heat-convection fields (in Japanese). *Trans. Jpn. Soc. Mech. Eng. B* **73**(733), 1884–1891 (2007)
91. E. Katamine, N. Nishimura, H. Azegami, Shape optimization of steady-state viscous flow fields for drag minimization and lift maximization (in Japanese). *Trans. Jpn. Soc. Mech. Eng. B* **74**(748), 2426–2434 (2008)

92. E. Katamine, T. Tsubata, H. Azegami, Solution to shape optimization problem of viscous flow fields considering convection term, in *Inverse Problems in Engineering Mechanics IV*, ed. by M. Tanaka (Elsevier, Tokyo, 2003), pp. 401–408
93. E. Katamine, H. Yoshioka, K. Matsuura, H. Azegami, Shape optimization of thermoelastic fields for mean compliance minimization (in Japanese). *Trans. Jpn. Soc. Mech. Eng. C* **77**(783), 4015–4023 (2011)
94. A. Kawamoto, T. Matsumori, S. Yamasaki, T. Nomura, T. Kondoh, S. Nishiwaki, Heaviside projection based topology optimization by a PDE-filtered scalar function. *Struct. Multidiscip. Optim.* **44**, 19–24 (2011)
95. F. Kikuchi, *Mathematics of Finite Element Method: Mathematical Basics and Error Analysis (in Japanese)* (Baifukan, Tokyo, 1994)
96. F. Kikuchi, *Overview of Finite Element Method, New and Revised Edition (in Japanese)* (Saiensu-sha, Tokyo, 1999)
97. M. Kimura, Shape derivative of minimum potential energy: abstract theory and applications, in *Jindřich Nečas Center for Mathematical Modeling Lecture Notes Volume IV, Topics in Mathematical Modeling*, pp. 1–38 (2008)
98. M. Kimura, I. Wakano, New mathematical approach to the energy release rate in crack extension (in Japanese). *Trans. Jpn. Soc. Ind. Appl. Math.* **16**(3), 345–358 (2006)
99. R.V. Kohn, R. Lipton, Optimal bounds for the effective energy of a mixture of isotropic, incompressible, elastic materials. *Arch. Ration. Mech. Anal.* **102**(4), 331–350 (1988)
100. R.V. Kohn, G. Strang, Optimal design and relaxation of variational problems, part 1. *Commun. Pure Appl. Math.* **39**, 1–25 (1986)
101. R.V. Kohn, G. Strang, Optimal design and relaxation of variational problems, part 2. *Commun. Pure Appl. Math.* **39**, 139–182 (1986)
102. R.V. Kohn, G. Strang, Optimal design and relaxation of variational problems, part 3. *Commun. Pure Appl. Math.* **39**, 353–377 (1986)
103. R.V. Kohn, G. Strang, Optimal design in elasticity and plasticity. *Int. J. Numer. Methods Eng.* **22**, 183–188 (1986)
104. E. Kreyszig, *Introductory Functional Analysis with Applications* (Wiley, New York, 1978)
105. S. Kurodai, *Functional Analysis (in Japanese)* (Kyoritsu Shuppan, Tokyo, 1980)
106. S. Kushimoto, *Basics of Optimization Problems (in Japanese)* (Morikita Publishing, Tokyo, 1979)
107. B.S. Lazarov, O. Sigmund, Filters in topology optimization based on Helmholtz-type differential equations. *Int. J. Numer. Methods Eng.* **86**(6), 765–781 (2011)
108. R.S. Lehman, Developments at an analytic corner of solutions of elliptic partial differential equations. *J. Appl. Math. Mech.* **8**, 727–760 (1959)
109. J.L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*. Translated by S.K. Mitter (Springer, Berlin, 1971)
110. K.A. Lurie, A.V. Cherkaev, A.V. Fedorov, Regularization of optimal design problems for bars and plates, part 1. *J. Optim. Theory Appl.* **37**, 499–522 (1982)
111. K.A. Lurie, A.V. Cherkaev, A.V. Fedorov, Regularization of optimal design problems for bars and plates, part 2. *J. Optim. Theory Appl.* **37**, 523–543 (1982)
112. Z.D. Ma, N. Kikuchi, I. Hagiwara, Structural topology and shape optimization for a frequency response problem. *Computational Mechanics* **13**, 157–174 (1993)
113. K. Matsui, K. Terada, Continuous approximation of material distribution for topology optimization. *Int. J. Numer. Methods Eng.* **59**, 1925–1944 (2004)
114. W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations* (Cambridge University Press, Cambridge, 2000)
115. G.H. Meisters, C. Olech, Locally one-to-one mappings and a classical theorem on schlicht functions. *Duke Math. J.* **30**, 63–80 (1963)
116. S. Miyajima, *Basics and Applications of Sobolev Spaces (in Japanese)* (Kyoritsu Shuppan, Tokyo, 2006)
117. H.P. Mlejnek, R. Schirrmacher, An engineer's approach to optimal material distribution and shape finding. *Comput. Methods Appl. Mech. Eng.* **106**, 1–26 (1993)

118. B. Mohammadi, O. Pironneau, *Applied Shape Optimization for Fluids* (Oxford University Press, Oxford, New York, 2001)
119. B. Mohammadi, O. Pironneau, *Applied Shape Optimization for Fluids*, 2nd edn. (Oxford University Press, Oxford, New York, 2010)
120. D. Murai, *Error analyses in numerical solutions for initial and boundary value problems of partial differential equations and shape optimization problems (in Japanese)*, PhD thesis, Nagoya University, 2011
121. D. Murai, H. Azegami, Error analysis of H1 gradient method for topology optimization problems of continua. *JSIAM Lett.* **3**, 73–76 (2011)
122. D. Murai, H. Azegami, Error analysis of H1 gradient method for shape-optimization problems of continua. *JSIAM Lett.* **5**, 29–32 (2013)
123. F. Murat, Contre-exemples pour divers problèmes où le contrôle intervient dans les coefficients. *Ann. Mat. Pura Appl.* **4** *112*, 49–68 (1977)
124. F. Murat, S. Simon, Etudes de problèmes d'optimal design, in *Lecture Notes in Computer Science*, vol. 41 (Springer, Berlin, 1976), pp. 54–62
125. C.D. Murray, The physiological principle of minimum work applied to the angle of branching of arteries. *J. Gen. Physiol.* **9**, 835–841 (1926)
126. C.D. Murray, The physiological principle of minimum work. I. the vascular system and the cost of blood volume. *Proc. Natl. Acad. Sci.* **12**, 207–214 (1926)
127. K.G. Murty, *Linear Complementarity, Linear and Nonlinear Programming* (Heldermann Verlag, Berlin, 1988)
128. S. Nishiwaki, K. Izui, N. Kikuchi, *Topology Optimization (in Japanese)* (Maruzen, Tokyo, 2013)
129. K. Ohtsuka, A. Khudnev, Generalized J-integral method for sensitivity analysis of static shape design. *Control Cybern.* **29**, 513–533 (2000)
130. K. Ohtsuka, T. Takaishi, *Finite Element Analysis Using Mathematical Programming Language FreeFEM++ (in Japanese)* (Kyoritsu Shuppan, Tokyo, 2014)
131. L.H. Olesen, F. Okkels, H. Bruus, A high-level programming-language implementation of topology optimization applied to steady-state Navier-Stokes flow. *Int. J. Numer. Methods Eng.* **65**, 975–1001 (2006)
132. O. Pironneau, On optimum profiles in Stokes flow. *J. Fluid Mech.* **59**(1), 117–128 (1973)
133. O. Pironneau, On optimum design in fluid mechanics. *J. Fluid Mech.* **64**(1), 97–110 (1974)
134. O. Pironneau, *Optimal Shape Design for Elliptic Systems* (Springer, New York, 1984)
135. G. Polya, Torsion rigidity, principal frequency, electrostatic capacity and symmetrization. *Q. Appl. Math.* **6**, 267–277 (1948)
136. I. Raasch, M.S. Chargin, R. Bruns, Optimierung von pkw-bauteilen in bezug auf form und dimensionierung. *VDI Berichte* (699), 713–748 (1988)
137. S.F. Rahmatalla, C.C. Swan, A Q4/Q4 continuum structural topology optimization implementation. *Struct. Multidiscip. Optim.* **27**, 130–135 (2004)
138. G.I.N. Rozvany, M. Zhou, T. Birker, Generalized shape optimization without homogenization. *Structural Optimization* **4**, 250–254 (1992)
139. W. Rudin, *Principles of Mathematical Analysis* (McGraw-Hill, New York, 1976)
140. K. Shiga, *Flying to the Infinite: The Birth of Set Theory (in Japanese)* (Kinokuniya, Tokyo, 2008)
141. M. Shimoda, H. Azegami, H. Ihara, T. Sakurai, Shape optimization of linear elastic structures subject to multiple loading conditions: A traction method approach to minimum volume design (in Japanese). *Trans. JSME A* **61**(587), 1545–1552 (1995)
142. M. Shimoda, H. Azegami, T. Sakurai, Multiobjective shape optimization of linear elastic structures considering multiple loading conditions: Dealing with mean compliance minimization problems (in Japanese). *Trans. JSME A* **61**(582), 359–366 (1995)
143. M. Shimoda, H. Azegami, T. Sakurai, Boundary shape determination of continua with desired stress distribution (in Japanese). *Trans. JSME A* **62**(602), 2393–2400 (1996)
144. M. Shimoda, H. Azegami, T. Sakurai, Shape determination of continua for homologous deformation (in Japanese). *Trans. JSME A* **62**(604), 2831–2837 (1996)

145. M. Shimoda, H. Azegami, T. Sakurai, Numerical solution for min-max problems in shape optimization: Minimum design of maximum stress and displacement (in Japanese). *Trans. JSME A* **63**(607), 610–617 (1997)
146. M. Shimoda, J. Tsuji, H. Azegami, Minimum weight shape design for the natural vibration problem of plate and shell structures, in *Computer Aided Optimum Design in Engineering IX*, ed. by S. Hernandez, C.A. Brebbia (WIT Press, Southampton, UK, 2005), pp. 147–156
147. M. Shimoda, J. Tsuji, H. Azegami, Non-parametric shape optimization method for thin-walled structures under strength criterion, in *Computer Aided Optimum Design in Engineering X*, ed. by S. Hernandez, C.A. Brebbia (WIT Press, Southampton, UK, 2007), pp. 179–188
148. M. Shimoda, Z.C. Wu, H. Azegami, T. Sakurai, Numerical method for domain optimization problems using a general purpose FEM code: Traction method approach (in Japanese). *Trans. JSME A* **60**(578), 2418–2425 (1994)
149. K. Shintani, T. Nagatani, S. Ito, H. Azegami, Examination of shape optimization for the nonlinear buckling phenomenon of suspension parts (in Japanese). *Trans. JSME A* **74**(748), 1187–1198 (2011)
150. O. Sigmund, K. Maute, Topology optimization approaches: A comparative review. *Struct. Multidiscip. Optim.* **48**, 1031–1055 (2013)
151. O. Sigmund, J. Petersson, Numerical instabilities in topology optimization: A survey on procedures dealing with checkerboards, mesh-dependencies and local minima. *Structural Optimization* **16**, 68–75 (1998)
152. J. Simon, Differentiation with respect to the domain in boundary value problems. *Numer. Funct. Anal. Optim.* **2**(7–8), 649–687 (1980)
153. J. Simon, Second variations for domain optimization problems. *Control Theory Distrib. Parameter Syst. Appl.* **91**, 361–378 (1989)
154. J. Sokolowski, J.P. Zolésio, *Introduction to Shape Optimization: Shape Sensitivity Analysis* (Springer, New York, 1992)
155. G. Strang, G.J. Fix, *An Analysis of the Finite Element Method* (Prentice-Hall, Englewood Cliffs, NJ, 1973)
156. K. Suzuki, N. Kikuchi, A homogenization method for shape and topology optimization. *Comput. Methods Appl. Mech. Eng.* **93**, 291–318 (1991)
157. M. Tabata, in *Numerical Solutions of Differential Equations II* (in Japanese). Iwanami Kouza Applied Mathematics (Iwanami Shoten, Tokyo, 1994)
158. M. Tabata, in *Numerical Analysis of Partial Differential Equations* (in Japanese). Iwanami Kouza Applied Mathematics (Iwanami Shoten, Tokyo, 2010)
159. M. Tabata, H. Fujii, T. Miyoshi, Finite element method using singular function (in Japanese). *Bit* **5**, 1035–1040 (1973)
160. T. Tago, T. Aoki, H. Azegami, Identification of building damage using vibrational eigenvalue and eigenmode pairs. *Jpn. J. Ind. Appl. Math.* **32**(2), 297–313 (2015)
161. T. Takagi, *Analysis Overview, 3rd Rev.* (in Japanese) (Iwanami Shoten, Tokyo, 1961)
162. A. Tamura, M. Matsumura, *Optimization Method* (in Japanese) (Kyoritsu Shuppan, Tokyo, 2002)
163. U. Tautenhahn, On the asymptotical regularization of nonlinear ill-posed problems. *Inverse Problems* **10**, 1405–1418 (1994)
164. J.E. Taylor, M.P. Bendsøe, An interpretation for min-max structural design problems including a method for relaxing constraints. *Int. J. Solids Struct.* **20**, 301–314 (1984)
165. J.A. Trangenstein, *Numerical Solution of Elliptic and Parabolic Partial Differential Equations* (Cambridge University Press, Cambridge, 2013)
166. G.N. Vanderplaats, *Numerical Optimization Techniques for Engineering Design: With Applications* (McGraw-Hill, New York, 1984)
167. G.N. Vanderplaats, H. Miura, GENESIS-structural synthesis software using advanced approximation techniques. AIAA Report (92-4839-CP), pp. 180–190 (1992)
168. F. Wang, B.S. Lazarov, O. Sigmund, On projection methods, convergence and robust formulations in topology optimization. *Struct. Multidiscip. Optim.* **43**, 767–784 (2011)

169. M.Y. Wang, X. Wang, D. Guo, A level set method for structural topology optimization. *Comput. Methods Appl. Mech. Eng.* **192**, 227–246 (2003)
170. F. Warner, *Foundations of Differentiable Manifolds and Lie Groups* (Springer, New York, 1983)
171. K. Washizu, *Introduction to Energy Principles (in Japanese)* (Baifukan, Tokyo, 1980)
172. P. Wolfe, Convergence conditions for ascent methods. *SIAM Rev.* **11**, 226–235 (1969)
173. Z. Wu, H. Azegami, Domain optimization analysis in frequency response problems: Approach using traction method (in Japanese). *Trans. Jpn. Soc. Mech. Eng. C* **61**(590), 3968–3975 (1995)
174. Z. Wu, H. Azegami, Domain optimization analysis in natural vibration problems: mass minimization problems (in Japanese). *Trans. Jpn. Soc. Mech. Eng. C* **61**(587), 2653–2696 (1995)
175. Z. Wu, Y. Sogabe, H. Azegami, Domain optimization analysis in frequency response problems considering proportional viscous damping (in Japanese). *Trans. Jpn. Soc. Mech. Eng. C* **64**(623), 2618–2624 (1998)
176. Z.C. Wu, H. Azegami, Domain optimization analysis in natural vibration problems: approach using traction method (in Japanese). *Trans. Jpn. Soc. Mech. Eng. C* **61**(583), 930–937 (1995)
177. Z.Q. Wu, Y. Sogabe, Y. Arimitsu, H. Azegami, Shape optimization of transient response problems, in *Inverse Problems in Engineering Mechanics II*, ed. by M. Tanaka, G.S. Dulikravich (Elsevier, Tokyo, 2000), pp. 285–294
178. Z.Q. Wu, Y. Sogabe, H. Azegami, Shape optimization analysis for frequency response problems of solids with proportional viscous damping. *Key Eng. Mater.* **145–149**, 227–232 (1997)
179. H. Yabe, *Engineering Basics: Optimization and Its Applications (in Japanese)* (Suurikougaku-sha, Tokyo, 2006)
180. T. Yamada, K. Izui, S. Nishiwaki, A. Takezawa, A topology optimization method based on the level set method incorporating a fictitious interface energy. *Comput. Methods Appl. Mech. Eng.* **199**, 2876–2891 (2010)
181. H. Yamakawa, 53 others, *Optimum Design Handbook: Basics, Strategy and Application (in Japanese)* (Asakura Publishing, Tokyo, 2003)
182. R.J. Yang, C.H. Chuang, Optimal topology design using linear programming. *Comput. Struct.* **52**, 265–275 (1994)
183. Z. Yoshida, *New Publication Functional Analysis for Applications: Its Way of Thinking and Techniques (in Japanese)* (Saiensu-sha, Tokyo, 2006)
184. K. Yosida, *Functional Analysis*, 6th edn. (Springer, Berlin; New York, 1980)
185. J. P. Zolésio, Domain variational formulation for free boundary problems, in *Optimization of Distributed Parameter Structures*, vol. 2, ed. by E.J. Haug, J. Cea, (Sijthoff & Noordhoff, Alphen aan den Rijn, The Netherlands, 1981), pp.1152–1194
186. J.P. Zolésio, The material derivative (or speed) method for shape optimization, in *Optimization of Distributed Parameter Structures*, vol. 2, ed. by E.J. Haug, J. Cea (Sijthoff & Noordhoff, Alphen aan den Rijn, The Netherlands, 1981), pp. 1089–1151

Index

A

Abstract minimization problem, 233
Abstract saddle point problem, 251, 254
Abstract space, 171
Abstract variational problem, 228
Action integral, 161
Active, 22
Active set method, 134, 352, 355, 389
Addition, 171
Adjoint problem, 16, 28, 77
Adjoint space, 201
Adjoint variable, 8, 16, 34, 77, 96
Adjoint variable method, 15, 27, 78
Admissible set, 66
— of adjoint variables, 370, 478
— of design variables, 9, 46, 332, 365, 434
— of state variables, 334, 365, 465
Affine mapping, 305, 320
Affine subspace, 175
Almost everywhere, 183
Area coordinates, 303
Armijo criterion, 117, 353
Ascoli–Arzelà theorem, 373
Augmented function method, 131

B

Banach perturbation theory, 126
Banach space, 179
 reflexive —, 203
Barrier method, 131
Basis function, 261, 267
 — in finite element, 276, 288
 — in one-dimensional finite element
 method, 274

— in two-dimensional finite element
 method, 287
Basis vector method, 428
 β method, 425
Bijection, 575
Bilinear form, 200
 bounded —, 200
Bilinear operator, 200
 bounded —, 200
Bisection method, 114
Boolean matrix, 277, 290
Boundary, 567
Boundary condition, 579
 first-type —, 226, 579
 fundamental —, 226
 natural —, 226, 579, 582
 second-type —, 226, 579
 third-type —, 227
Boundary point, 567
Boundary value problem, 577
 mixed —, 226
Bounded, 228
Bounded set, 567
Broyden–Fletcher–Goldfarb–Shanno method, 128
Broyden method, 128
Bulk modulus, 245

C

Calculus of variations, 162
Calderón extension theorem, 199, 433
Cantor’s diagonal argument, 178
Cardinality, 178
Cauchy–Riemann equations, 237

Cauchy sequence, 177
 Cea's lemma, 272
 Chain rule of differentiation, 13
 Characteristic function, 359
 Chebyshev norm, 180
 Closed set, 567
 Closure, 567
 Coefficient matrix, 264, 278, 291
 total —, 281, 294
 Coefficient of viscosity, 249, 542
 Coercive, 165, 228
 Compact, 178
 Compactly embedded, 207
 Complementarity condition, 22, 89
 Complete, 177
 Completed set, 187
 Completion, 178
 Concave function, 56
 Conjugate, 123
 Connected, 573
 Constitutive equation, 6, 243
 Constitutive law, 6, 243
 Constraint function, 2, 9
 equality —, 46
 inequality —, 46
 Continuity equation, 249, 257
 Continuous, 567
 Convergence order, 107
 Convex function, 56
 Convex optimization problem, 49, 56
 Convex programming problem, 49
 Cost function, 2, 9, 33
 Cottle's constraint qualification, 87
 Cross-sectional-area gradient, 11, 34
 Cross-sectional derivatives, 10, 34
 second-order —, 11

D

Darcy law, 410
 Davidon–Fletcher–Powell method, 128
 Dense, 177
 Derivative, 53
 second-order —, 53
 total —, 53
 Descent angle, 110
 Descent direction, 107
 Design variable, 2, 9, 33
 Diameter, 319
 Dirac delta function, 189
 Dirac distribution, 189
 Direct differentiation method, 13, 26, 77
 Direct method, 107
 Dirichlet boundary, 226

Dirichlet condition, 226, 579
 homogeneous —, 175, 199
 inhomogeneous —, 175
 Dirichlet problem, 226
 Discrete optimization problem, 51
 Discrete programming problem, 51, 56
 Displacement, 240
 Dissipation energy, 33
 Dissipation energy density, 39
 Distribution, 188
 partial derivative of —, 189
 Divergence free Hilbert space, 252
 Divergence theorem, 585
 Domain, 573
 Dual cone, 83, 86, 571
 Dual interior point method, 49
 Duality index, 204
 Duality theorem, 98
 Dual plane, 68
 Dual product, 201
 Dual set, 68
 Dual space, 201
 second —, 203
 Dual weak Cauchy sequence, 203
 Dual weak compact, 203
 Dual weak complete, 203
 Dual weak convergence, 203

E

Eigen equation, 607
 Eigenfrequency, 607
 Eigenmode, 607
 Eigenpair, 607
 Eigenvalue problem, 607
 Elastic potential energy density, 6
 Element node vector, 277
 Elliptic, 228, 244
 Ellipticity, 246
 Equation of heat conduction, 577, 579
 Equilibrium equation of force, 163, 245
 Essentially bounded, 185
 Essential supremum, 186
 Euclidean norm, 180
 Euclid space, 2
 Euler description, 436
 Extension operator, 199
 External work, 8

F

Farkas's lemma, 571
 Feasible direction set, 67, 83
 Feasible set, 46

Finite element, 273
 — affine-equivalent —, 320
 — division, 319
 first-order rectangular —, 307
 four-node isoparametric —, 312
 hexahedral —, 310
 isoparametric —, 311
 Lagrange family rectangular —, 307
 one-dimensional —, 273
 one-dimensional m -th order —, 301
 regular —, 319, 391, 505
 serendipity group rectangular —, 308
 — solution, 318
 tetrahedron —, 309
 triangular —, 286
 — triangular m -th order —, 305
 First order constraint qualification, 87
 First variation, 162
 Fixed in space, 462
 Fixed with material, 461
 Fletcher–Reeves formula, 124
 Flow velocity, 249
 Fourier’s law of heat conduction, 578
 Fréchet derivative, 211
 Friedrichs mollifier, 187
 Functional, 201
 Function space, 183
 Fundamental boundary condition, 579, 582

G

Gâteaux derivative, 209
 Galerkin method, 261
 Gauss–Green theorem, 585
 Gaussian node, 314
 Gaussian quadrature, 313, 316
 Gauss’s divergence theorem, 585
 Generalized J integral, 565
 Global convergence, 107, 122
 Gradient, 53, 212
 Gradient method, 109
 — abstract —, 347
 — conjugate —, 123
 — for constrained problems, 132

H

H^1 gradient method, 348
 — of domain variation type, 494
 — of θ -type, 383
 H^1 Newton method, 349
 — of domain variation type, 501
 — of θ -type, 387
 Hölder continuous, 183

Hölder index, 183
 Hölder’s inequality, 186, 586
 Hölder space, 184
 Hamilton function, 169
 Hamiltonian, 4
 Hamilton’s principle, 160
 Harmonic operator, 582
 Heaviside step function, 189
 Hermitian symmetry, 182
 Hesse form, 212
 Hesse gradient, 21, 37, 80, 129, 149, 153, 346, 349, 355, 382, 388, 391, 405, 419, 490, 501, 504, 530, 553
 Hesse matrix, 11, 53
 Hessian, 11, 344
 Hilbert space, 182
 Homogeneous type, 226
 Hooke’s law, 6
 — generalized —, 244
 Hourglass mode, 318

I

Image cone, 571
 Image space, 68, 340, 570
 Implicit function theorem, 69, 339, 572
 Inactive, 22
 Incompressible fluid, 249
 Infinite sequence of points, 177
 Inf-sup condition, 253
 Inhomogeneous type, 226
 Initial condition, 579, 582
 Initial point, 107
 Inner point, 567
 Inner point method, 131
 Inner product, 182
 Inner product space, 182
 Interior, 567
 Interpolation error, 323
 Interpolation function, 276, 322
 Interpolation operator, 322
 Irregularity, 234
 Isomorphism, 197
 Iterative method, 106

J

Jacobian, 305, 612
 Jacobi determinant, 440
 Jacobi matrix, 67, 305, 440

K

Karush–Kuhn–Tucker conditions, 21, 38, 89, 132

Kernel space, 68, 570
 Kinetic energy, 160
 Known term vector, 264, 280, 292
 total —, 281, 294
 Korn's inequality, 588
 Korn's second inequality, 248, 588

L

Lagrange basis polynomials, 316
 Lagrange function, 22
 — for cost function, 16
 — in mechanics, 160
 — for an optimization problem, 71
 — for state determination problem, 8, 334
 Lagrange interpolation, 316
 Lagrange multiplier, 8, 16, 22, 34, 71
 Lagrange multiplier method, 15, 342
 — for an optimization problem with an
 equality constraint, 71
 — for an optimization problem with an
 inequality constraint, 89

Lagrangian description, 436
 Lamé's parameters, 244
 Laplace equation, 224
 Laplace operator, 224, 582
 Laplace problem, 224
 Lax–Milgram theorem, 229
 Lebesgue integrable, 185
 Lebesgue integral, 185
 Lebesgue measure, 183
 Lebesgue's convergence theorem, 186
 Lebesgue space, 185

Legendre polynomials, 314
 Length coordinate, 299
 Linear combination, 171
 Linear elastic problem, 246
 Linear form, 197
 Linear functional, 201
 bounded —, 201
 Linear hull, 175
 Linearized feasible direction set, 86
 Linear mapping, 197
 Linear operation, 171
 Linear operator, 197
 bounded —, 197
 continuous —, 197
 Linear optimization problems, 49
 Linear programming problems, 49
 — of design variables, 9, 364, 433
 — of state variables, 9, 365, 465
 Linear space, 171
 Linear span, 175
 Lipschitz boundary, 172, 573

Lipschitz constant, 183
 Lipschitz continuous, 184
 Lipschitz domain, 172, 183, 223, 574
 Lipschitz space, 184
 Local minimum point, 51
 Local minimum value, 51
 Local node number, 277, 288
 Longitudinal elastic modulus, 6, 245
 Lower semi-continuity, 52
 Lower semi-continuous, 337, 568

M

Material derivative, 436
 Mathematical programming, 105
 Maximum descent method, 110
 Maximum norm, 180
 Mean compliance, 8, 399, 520
 Mean curvature, 576
 Mean flow resistance, 33, 412, 544
 Mean value theorem, 54
 Metric, 176
 Metric space, 176, 567
 Minimum point, 51
 Minimum principle of potential energy, 163
 Minimum value, 51
 Minkowski's inequality, 186, 587
 Motion equation, 160, 162
 Multi-index, 172
 Multi-objective optimization problem, 50
 Multiply connected, 573

N

Navier–Stokes equation, 257
 Navier–Stokes problem, 257
 Negative definite, 58
 semi- —, 58
 Neighborhood, 567
 Neumann boundary, 226
 Neumann condition, 226, 579
 Neumann problem, 226
 Newton method, 125
 abstract —, 348
 — for constrained problems, 146
 quasi- —, 128
 Newton–Raphson method, 128
 Newton viscosity, 249
 Nodal value vector, 275, 287
 Dirichlet-type —, 275, 287
 element —, 276, 288
 Neumann-type —, 275, 288
 total —, 281, 294

- Node, 273, 287
 element —, 288
 Non-linear optimization problem, 49
 Non-linear programming problem, 49
 Non-positive cone, 571
 Norm, 179
 semi- —, 184, 192
 Normal, 576
 Normal element, 299
 Normed space, 179
 Null space, 68, 570
- O**
 Objective function, 2, 9, 46
 Open set, 567
 Operations research, 105
 Operator, 197
 Optimality condition, 2, 21
 Order estimation of error, 324
 Orthogonal complement space, 570
 Outer point method, 131
 Outer unit normal, 576
- P**
 Pareto solution, 50
 Partial differential equation, 583
 elliptic —, 246, 580, 583
 hyperbolic —, 583
 parabolic —, 580, 583
 second-order —, 583
 Partial shape derivative of a function, 436
 Penalty method, 131
 Piecewise C^k class boundary, 575
 Piola transformation, 442
 p -norm, 180
 Poincaré's inequality, 230, 587
 Point, 171
 Poisson equation, 224
 homogeneous —, 224
 Poisson problem, 224
 Poisson's ratio, 245
 Polak–Ribiére formula, 124
 Pontryagin's local minimum condition, 169
 Pontryagin's minimum principle, 169
 Positive definite, 13, 58
 — real symmetric matrix, 58, 569
 semi- —, 58
 Potential energy, 3, 7, 160
 elastic —, 163
 external —, 4, 163
 — of external force, 163
 internal —, 4, 163
- Power loss, 33
 Pressure, 249
 Principal curvature, 577
 Principle of virtual work, 8
 Pyramidal function, 287
- Q**
 Quadratic optimization problem, 49
 Quadratic programming problem, 49
- R**
 Range space, 68, 570
 Rank, 67
 Rank d material, 360
 Real linear space, 171
 Regularity, 234
 Relative compact, 178
 Rellich–Kondrachov compact
 embedding theorem, 207
 Riesz's representation theorem, 208, 229
 Right inverse operator, 198
 Ritz method, 270
 Robin condition, 227
 Robin problem, 227
 Rolle's theorem, 54
 Rotation tensor, 241
- S**
 Saddle point theorem, 98
 Scalar, 171
 Scalar product, 171, 182
 Schwartz distribution, 188
 partial derivative of —, 189
 Schwarz's inequality, 587
 Search direction, 106
 Search vector, 106
 Secant method, 114, 158
 Second variation, 162
 Selected reduced integration, 318
 Self-adjoint relationship, 17, 34, 401, 414, 521,
 546
 Separable, 177
 Shape derivative, 435
 — of a function, 435
 — of a functional, 438
 second-order —, 439
 Shape function, 276
 Shape gradient, 438
 Shape Hessian, 439
 Shear modulus, 244
 Side constraint, 9, 333, 365, 373, 476

Sigmoid functions, 364
 Simplex method, 49
 Simply connected, 573
 Slack variable, 89
 Slater constraint qualification, 98
 Sobolev embedding theorem, 193
 Sobolev inequality, 194
 Sobolev space, 190
 Solid isotropic material with penalization (SIMP), 361
 Spatial derivative, 436
 State determination problem, 2, 3, 366
 State equation, 2
 State variable, 2, 9, 33
 Stationary point, 63
 Step size, 106, 114
 Stokes equation, 249
 Stokes problem, 249
 Strain, 6, 241
 linear —, 241
 Strain energy density, 6, 23, 244, 533
 Stress, 6, 242
 Cauchy —, 242, 250
 Strict line search method, 114
 Strong convergence, 202
 Strong form, 226
 Super p -th order convergence, 107
 Support, 172
 Sylvester's criterion, 569

T

Tangent, 576
 Tangential cone, 83
 Tangential plane, 67
 Taylor expansion, 55
 Taylor's theorem, 54
 Terminal condition of velocity, 162
 Test function, 188
 θ -derivative, 374
 Topology optimization problem, 359
 — of density type, 361
 — of θ -type, 362, 369
 Trace operator, 198, 509

Trace theorem, 198
 Traction, 163, 240
 Traction method, 429, 498
 — of Robin type, 499
 — with spring, 499
 Trial point, 107
 Triangle inequality, 587

U
 Uncertainty by constant, 232
 Uniformly continuous, 568

V
 Varying with boundary measure, 463
 Varying with domain measure, 463
 Vector, 171
 Vector space, 171
 Volume coordinates, 309
 Volume force, 163, 240, 249

W

Weak Cauchy sequence, 202
 Weak compact, 202
 Weak complete, 202
 Weak convergence, 202
 Weak form, 226
 — of linear elastic problem, 247
 — of Poisson problem, 225
 — of Stokes problem, 251
 Weak solution, 226
 Weierstrass's approximation theorem, 180
 Weierstrass's theorem, 52, 335
 Wolfe criterion, 119, 353

Y

Young's modulus, 6, 162, 245

Z

Zoutendijk condition, 123