Advances in Mechanics and Mathematics 36

Zdeněk Dostál
Tomáš Kozubek
Marie Sadowská
Vít Vondrák

Scalable Algorithms for Contact Problems

Springer

Advances in Mechanics and Mathematics

Volume 36

More information about this series at http://www.springer.com/series/5613

Zdeněk Dostál · Tomáš Kozubek
Marie Sadowská · Vít Vondrák

Scalable Algorithms for Contact Problems

Springer

Zdeněk Dostál
National Supercomputer Center and
Department of Applied Mathematics
VŠB-Technical University of Ostrava
Ostrava
Czech Republic

Tomáš Kozubek
National Supercomputer Center and
Department of Applied Mathematics
VŠB-Technical University of Ostrava
Ostrava
Czech Republic

Marie Sadowská
Department of Applied Mathematics
VŠB-Technical University of Ostrava
Ostrava
Czech Republic

Vít Vondrák
National Supercomputer Center and
Department of Applied Mathematics
VŠB-Technical University of Ostrava
Ostrava
Czech Republic

ISSN 1571-8689 ISSN 1876-9896 (electronic)
Advances in Mechanics and Mathematics
ISBN 978-1-4939-6832-9 ISBN 978-1-4939-6834-3 (eBook)
DOI 10.1007/978-1-4939-6834-3

Library of Congress Control Number: 2016958490

Mathematics Subject Classification (2010): 35Q74, 65K15, 65N55

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Science+Business Media LLC
The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

To our families

Preface

The practical interest in contact problems stems from the fundamental role of contact in mechanics of solids and structures. Indeed, the contact of one body with another is a typical way how loads are delivered to a structure and the typical mechanism which supports structures to sustain the loads. Thus we do not exaggerate much if we say that the contact problems are in the heart of mechanical engineering.

The contact problems are also interesting from the computational point of view. The conditions of equilibrium of a system of bodies in mutual contact enhance a priori unknown boundary conditions, which make the contact problems strongly nonlinear, and when some of the bodies are "floating," then the boundary conditions admit rigid body motions and a solution need not be unique. After the discretization, the contact problem can be reduced to a finite dimensional problem, such as the minimization of a possibly non-differentiable convex function in many variables (currently from tens of thousands to billions) subject to linear or nonlinear inequality constraints for surface variables, with a specific sparse structure. Due to the floating bodies, the cost function can have a positive semidefinite quadratic part. Thus the solution of large discretized multibody contact problems still remains a challenging task, which can hardly be solved by general algorithms.

The main purpose of this book is to present scalable algorithms for the solution of multibody contact problems of linear elasticity, including the problems with friction and dynamic contact problems. Most of these results were obtained during the last twenty years. Let us recall that an algorithm is said to be *numerically scalable* if the cost of the solution increases nearly proportionally to the number of unknown variables, and it enjoys *parallel scalability* if the computational time can be reduced nearly proportionally to the number of processors. The algorithms which enjoy numerical scalability are in a sense optimal as the cost of the solution by such algorithms increases as the cost of duplicating the solution. Taking into account the above characterization of contact problems, it is rather surprising that such algorithms exist.

Our development of scalable algorithms for contact problems is based on the following observations:

- There are algorithms which can solve relevant quadratic programming and QCQP problems with asymptotically linear complexity.
- Duality based methods like FETI let us define a sufficiently small linear subspace with the solution.
- The projector to the space of rigid body motions can be used to precondition both the linear and nonlinear steps of solution algorithms.
- The space decomposition used by the variants of FETI is an effective tool for solving multibody contact problems and opens a way to the massively parallel implementation of the algorithms.

The development of scalable algorithms represents a challenging task even when we consider much simpler linear problems. For example, the computational cost of solving a system of linear equations arising from the discretization of the conditions of equilibrium of an elastic body with prescribed displacements or traction on the boundary by direct sparse solvers increases typically with the square of the number of unknown nodal displacements. The first numerically scalable algorithms for linear problems of computational mechanics based on the concept of multigrid came in use only in the last quarter of the last century and fully scalable algorithms based on the FETI (Finite Element Tearing and Interconnecting) methods were introduced by Farhat and Roux by the end of the twentieth century.

The presentation of the algorithms in the book is complete in the sense that it starts from the formulation of contact problems, briefly describes their discretization and the properties of the discretized problems, provides the flowcharts of solution algorithms, and concludes with the analysis, numerical experiments, and implementation details. The book can thus serve as an introductory text for anybody interested in contact problems.

Synopsis of the book:

The book starts with a general introduction to contact problems of elasticity with the account of the main challenges posed by their numerical solution. The rest of the book is arranged into four parts, the first of which reviews some well-known facts on linear algebra, optimization, and analysis in the form that is useful in the following text.

The second part is concerned with the algorithms for minimizing a quadratic function subject to linear equality constraints and/or convex separable constraints. A unique feature of these algorithms is their rate of convergence and the error bounds in terms of bounds on the spectrum of the Hessian of the cost function. The description of the algorithms is organized in five chapters starting with a separate overview of two main ingredients, the conjugate gradient (CG) method (Chap. 5) for unconstrained optimization and the results on gradient projection (Chap. 6), in particular on the decrease of the cost function along the projected-gradient path.

Chapters 7 and 8 describe the MPGP (Modified Proportioning with Gradient Projections) algorithm for minimizing strictly convex quadratic functions subject to separable constraints and its adaptation MPRGP (Modified Proportioning with Reduced Gradient Projections) for bound constrained problems. The result on the rate of convergence in terms of bounds on the spectrum of the Hessian matrix is given that guarantees a kind of optimality of the algorithms—it implies that the number of iterates that are necessary to get an approximate solution of any instance of the class of problems with the spectrum of the Hessian contained in a given positive interval is uniformly bounded regardless the dimension of the problem. A special attention is paid to solving the problems with elliptic constraints and coping with their potentially strong curvature, which can occur in the solution of contact problems with orthotropic friction.

Chapter 9 combines the algorithms for solving problems with separable constraints and a variant of the augmented Lagrangian method in order to minimize a convex quadratic function subject to separable and equality constraints. The effective precision control of the solution of separable problems in the inner loop opened the way to the extension of the optimality results to the problems with separable and linear equality constraints that arise in the dual formulation of the conditions of equilibrium. Apart from the basic SMALSE (Semi-Monotonic Algorithm for Separable and Equality constraints) algorithm, the specialized variants for solving bound and equality constrained quadratic programming problems (SMALBE) and QCQP problems including elliptic constraints with strong curvature (SMALSE-Mw) are considered.

The most important results of the book are presented in the third part, including the scalable algorithms for solving multibody frictionless contact problems, contact problems with Tresca's friction, and transient contact problems.

Chapter 10 presents the basic ideas of the scalable algorithms in a simplified setting of multidomain scalar variational inequalities.

Chapters 11–13 develop the ideas presented in Chap. 10 to the solution of multibody frictionless contact problems, contact problems with friction, and transient contact problems. For simplicity, the presentation is based on the node-to-node discretization of contact conditions. The presentation includes the variational formulation of the conditions of equilibrium, the finite element discretization, some implementation details, the dual formulation, the TFETI (Total finite Element Tearing and Interconnecting) domain decomposition method, the proof of asymptotically linear complexity of the algorithms (numerical scalability), and numerical experiments.

Chapter 14 extends the results of Chaps. 10 and 11 to solving the problems discretized by the boundary element methods in the framework of the TBETI (Total Boundary Element Tearing and Interconnecting) method. The main new features include the reduction of the conditions of equilibrium to the boundary and the boundary variational formulation of the conditions of equilibrium.

Chapters 15 and 16 extend the results of Chaps. 10–14 to solving the problems with varying coefficients and/or with the non-penetration conditions implemented by mortars. It is shown that the reorthogonalization-based preconditioning or the

renormalization-based scaling can relieve the ill-conditioning of the stiffness matrices and that the application of the mortars need not spoil the scalability of the algorithms.

The last part begins with two chapters dealing with the extension of the optimality results to some applications, in particular to contact shape optimization and contact problems with plasticity. The book is completed by a chapter on massively parallel implementation and parallel scalability. The (weak) parallel scalability is demonstrated by solving an academic benchmark discretized by billions of nodal variables. However, the methods presented in the book can be used for solving much larger problems, as demonstrated by the results for a linear benchmark discretized by tens of billions of nodal variables.

Ostrava
January 2016

Zdeněk Dostál
Tomáš Kozubek
Marie Sadowská
Vít Vondrák

Acknowledgments

Many of the results presented in this book have been found by the authors in cooperation with the colleagues from other institutions. Here we would like to offer our thanks especially to Ana Friedlander and Mario Martínez for the early assessment of the efficiency of augmented Lagrangian methods, to Sandra A. Santos and F.A.M. Gomes for their share in the early development of algorithms for solving variational inequalities, to Olaf Steinbach for introducing us to the boundary element methods, to Joachim Schöberl for sharing his original insight into the gradient projection method, to Jaroslav Haslinger for joint research in the problems with friction, and to Charbel Farhat for drawing attention to the practical aspects of our algorithms and inspiration for thinking twice about simple topics.

Our thanks go also to our colleagues and students from the Faculty of Electrical Engineering and Computer Science of VŠB-Technical University of Ostrava. Jiří Bouchala offered his insight into the topics related to boundary integral equations, David Horák first implemented many variants of the algorithms that appear in this book, Marta Domorádová–Jarošová participated in the research of conjugate projectors, Lukáš Pospíšil participated in the development of algorithms for orthotropic friction and initiated the research in fast gradient methods, Petr Horyl, Alex Markopoulos, and Tomáš Brzobohatý paved the way from academic benchmarks to real-world problems, Petr Kovář helped with the graph theory, and Radek Kučera adapted the algorithms for bound constrained QP to the solution of more general problems with separable constraints and carried out a lot of joint work. Our special thanks go to Oldřich Vlach, who not only participated in the research related to mortars and composites with inclusions but also prepared many figures and participated in the experiments. The key figures in parallel implementation were Václav Hapla, David Horák, and Luboš Říha. The book would be much worse without critical reading of its parts by Dalibor Lukáš, Kristýna Motyčková, and other colleagues. We are also grateful to Renáta Plouharová for improving the English.

We gratefully acknowledge the support of IT4 Innovations Center of Excellence project, reg. no. CZ.1.05/1.1.00/02.0070 via 'Research and Development for Innovations' Operational Programme funded by the Structural Funds of the European Union.

Contents

Part III Scalable Algorithms for Contact Problems

Chapter 1
Contact Problems and Their Solution

We start our exposition by an informal presentation of contact problems, including those that motivated our research, with a brief discussion of the challenges arising in their numerical solution. Frictionless problems, problems with friction, and dynamic contact problems are considered. We discuss their specific features, especially those that complicate their solution or those that can be used to simplify the solution, recall some theoretical results concerning the existence and uniqueness of a solution, present basic ideas related to the development of scalable algorithms for the numerical solution of contact problems, and mention some historical results. In this chapter, we assume that the strains and displacements are small and within the elastic limit, at least in one time step in the case of transient problems, so that linear elasticity can be used to the formulation of the conditions of equilibrium.

1.1 Frictionless Contact Problems

We speak about frictionless contact problems whenever we can obtain an acceptable solution under the assumption that the tangential forces on the contact interface can be neglected. Many such problems arise in mechanical engineering whenever there is need to predict the stress or deformation of the moving parts of machines or vehicles.

The frictionless contact problems are the most simple ones, so it is not surprising that the first numerical results were obtained just for them. The first publication dates back to 1881 when Heinrich Hertz published his paper "On the contact of elastic solids" [1]. Hertz observed that when two smooth bodies come into contact so that the contact area is much smaller than the characteristic radius of each body, then the nonlinear non-penetration boundary conditions are confined to a small region of predictable shape and it is possible to simplify the conditions of equilibrium near the contact so that they can be solved analytically. A variant of such problem is depicted in Fig. 1.1.

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3_1

Hertz's results are still relevant for the design of bearings, gears, and other bodies when two smooth and nonconforming surfaces come into contact, the strains are small and within the elastic limit, the area of contact is much smaller than the characteristic radius of the bodies, and the friction can be neglected. We use them to verify the results of our benchmarks.

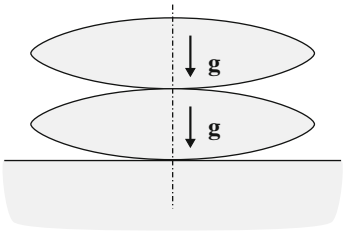

Fig. 1.1 Stacked lenses pressed together

As an example of the frictionless contact problem with a small contact interface that can not be solved analytically, let us consider the problem to describe the deformation and contact pressure in the ball bearings depicted in Fig. 1.2. A new feature of this problem is the complicated interaction of several bodies, some of them without prescribed displacements, through the nonlinear non-penetration conditions. The bodies are of different shapes – we can easily recognize balls, rings, and cages (see Fig. 1.3). The balls are not fixed in their cages, so their stiffness matrices are necessarily singular and the prescribed forces can not be arbitrary as the balls can be in equilibrium only if the moment of the external forces on the balls is equal to zero as in our case. Though the displacements and forces are typically given on the parts of the surfaces of some bodies, the exact places where the deformed balls come into contact with the cages or the rings are known only after the problem is solved. Moreover, the displacements of the balls would not be uniquely determined even if we replaced all the inequality constraints that describe the non-penetration by the corresponding equalities describing the related bilateral constraints.

Fig. 1.2 Ball bearings

Fig. 1.3 Ball bearings decomposition

A solution of the ball bearing and other frictionless contact problems with "floating" bodies exists provided the external forces acting on each such body are balanced (including the reaction forces). If the prescribed displacements of each body prevents its rigid body motion, then the solution exists and is necessary unique.

The first results concerning the existence and uniqueness of a solution to the contact problem date back to Gaetano Fichera, who published the paper "On the elastostatic problem of Signorini with ambiguous boundary conditions" [2] in 1964. Fichera coined the problem as Signorini's problem to honor his teacher, who draw his attention to it. Later presentation of the existence results based on coercivity of the energy function can be found, e.g., in Hlaváček, Haslinger, Nečas, and Lovíšek [3] or Kikuchi and Oden [4]. Let us point out that the most popular sufficient condition for the existence of a solution, the coercivity of the energy functional, does not hold for the bearing problems.

1.2 Contact Problems with Friction

The contact problems with friction arise whenever we have to take into account the tangential forces on contact interface. They are important in many areas of engineering including locomotive wheel–rail or tire–road contact, braking systems, yielding clamp connections, etc.

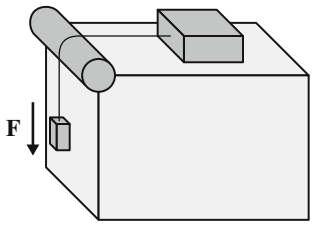

Fig. 1.4 Leonardo da Vinci experiment with friction

The effort to describe phenomenologically the friction forces dates back to Leonardo da Vinci. Carrying out the experiments like those depicted in Fig. 1.4, Leonardo observed that the area of contact interface has no effect on the friction and if the load on the object is doubled, the friction is also doubled. This is a special version of the most popular friction law used in many applications that is named after Charles-Augustin Coulomb. This law was formulated for quasistatic contact by Amontonos in 1691 and extended to dynamic situations by Coulomb in 1781 [5]. The Coulomb (also Amontonos–Coulomb) friction law claims that "strength due to friction is proportional to the compressive force." The Coulomb friction law agrees in many cases with the observation, but Coulomb himself observed that "for large bodies friction does not follow exactly this law."

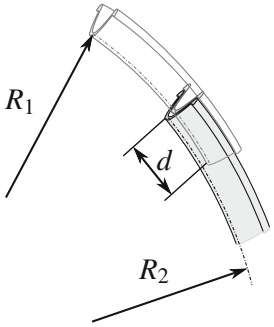

Fig. 1.5 Yielding support

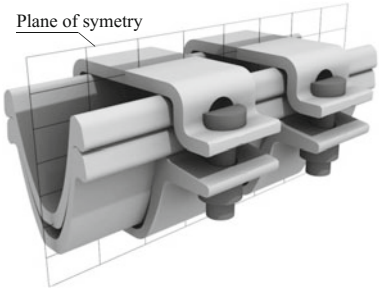

Fig. 1.6 Yielding clamp connection

A realistic example of a contact problem with friction is the analysis of the yield clamped connection of the mine support depicted in Figs. 1.5 and 1.6. The support comprises several parts that are deformed by applied forces, so it is not possible to assume that the bodies are rigid. As in the ball bearing example, some parts are also without prescribed displacements, so that the numerical algorithms for the solution must be prepared to deal with singular stiffness matrices of the discretized problem. A brand new difficulty is the necessity to balance the contact forces not only in the normal direction but also in the tangential direction. To solve the problem with friction, we have to identify the unknown slip–stick interface. The difficulties arise also in its variational formulation due to the non-differentiable dissipative term in the energy function and, in the case of Coulomb's friction, due to the lack of convexity.

Unlike the frictionless case, the existence theory for the quasistatic contact problems with Coulomb's friction is much weaker and does not guarantee that a solution exists for realistic problems. Moreover, a solution is known to be unique only in very exceptional cases. A recent overview of the results concerning the existence and uniqueness of a solution of contact problems with friction can be found in Eck, Jarůšek, and Krbec [6].

The most complete theory has been developed for the Tresca (given) friction, which assumes that the normal contact area and the pressure force, which defines the slip–stick bound, are known a priori. Though this assumption is not realistic, it is useful as the well-understood approximation of the Coulomb friction with strong theoretical results on the existence and uniqueness of a solution. In our book, we use the Tresca friction as an auxiliary problem in the fixed point algorithms for the Coulomb friction.

1.3 Transient Contact Problems

The transient contact problems arise whenever we have to take into account the inertia forces. It follows that it is physically more realistic to consider dynamic models than the static models considered above. Many such problems arise in mechanical engineering, geophysics, or biomechanics when the bodies in contact are moving fast, as in car engines.

However, to provide a useful solution to realistic problems requires not only to overcome the difficulties specified above for the static problems, but also to resolve the problems arising from the lack of smoothness of time derivatives, which puts high demand on the construction of effective time discretization schemes. Moreover, since the solution of transient problems is typically reduced to a sequence of related static problems, it is natural to assume that their solution is much more time consuming than the solution of related static problems.

Fig. 1.7 Crash test

An example of a realistic transient contact problem is the analysis of a crash test depicted in Fig. 1.7. A sophisticated time discretization is necessary to preserve the energy. A reasonable time discretization requires the solutions of 10^3–10^5 auxiliary static problems per second, so that an efficient solver of the static problems is necessary. Moreover, in each time step, it is necessary to identify a possible contact interface. The only simplification, as compared with the static problems, concerns the stiffness matrices of the "floating" bodies as they are regularized by the mass matrix.

In spite of a lot of research, a little is known about the solvability of the above problem, so a numerical solution is typically obtained under the assumption that it exists. Moreover, it is assumed that the solution is sufficiently smooth so that its second derivatives exist in some reasonable sense and can be approximated by finite differences. The complications arise from the hyperbolic character of the equations of motion.

1.4 Numerical Solution of Contact Problems

Given a system of elastic bodies in a reference configuration, traction, and boundary conditions, including contact conditions on a potential contact interface, the complete solution of the associated contact problem comprises the resulting displacements of the bodies and the reaction forces on the contact interface, from which we can evaluate the stress and strain fields in the bodies. To describe briefly how to resolve the challenges posed by the numerical solution of contact problems, we split the solution procedure into three stages:

- a continuous formulation of the conditions of equilibrium and the boundary conditions,
- the discretization,
- the solution of the discretized problem.

The choice of a *continuous formulation* is essential for the success of the next steps. Many effective solution methods are based on a variational formulation of the conditions of equilibrium, in particular on the observation that the equilibrium minimizes the energy function (considered as a function of the displacements) subject to the constraints specified by the boundary conditions. The energy function introduces a structure that can be exploited by the solution algorithms and simplifies the treatment of discontinuous coefficients and traction. The energy formulation in a suitable function space is also mathematically sound in the sense that it enables to obtain results on the existence and uniqueness of a solution. More general variational formulation of the equilibrium conditions is useful for the problems with additional nonlinearities, such as large deformations or plasticity.

If some additional assumptions are satisfied, it is possible to express the displacements in the interior of the bodies in terms of the displacements and their derivatives on the boundary. In this case, the energy function can be reduced to the boundary of the bodies so that it depends only on the displacements near the boundary, which reduces the dimension of the problem by one. Though the reduced variational formulation is not as general as the full one, there are problems which can be solved more efficiently in the *boundary formulation*, including the exterior problems, the problems with cracks, or the contact shape optimization problems. Moreover, the surface discretization is much simpler than the volume discretization.

The *discretization* of contact problems is partly the same as that of the linear problems of elasticity — the finite element method (FEM) and the boundary element method (BEM) are widely used for the discretization of the full and reduced energy functions, respectively. The discretization of contact conditions is more tricky. A straightforward linearization of the non-penetration conditions is possible and works for some simple problems, especially when a matching discretization is used. However, the matching discretization is hardly possible, e.g., for the transient contact problems or for the contact shape optimization problems as it would require remeshing which could affect related cost functions.

More sophisticated discretization is important also when the contact interface is large and curved as that of the hip joint substitute depicted in Fig. 1.8. The remedy consists in imposing the contact conditions by local averages. Strong approximation properties of the variationally *consistent mortar discretization* introduced by Wohlmuth can be found in [7].

Fig. 1.8 Hip joint substitute (*left*) and its model decomposed into subdomains (*right*)

Efficient algorithms for the *solution* of contact problems typically combine a fast solver of auxiliary linear problems with a strategy for effective identification of the constraints that are active in the solution. Recent development of direct sparse linear solvers extended their applicability to larger problems, currently millions of unknowns, but for the solution of still larger problems, it is necessary to use a suitable scalable solver that can currently find an approximate solution of the systems with billions of unknowns and exploit effectively some tens of thousands of processors.

The scalable solvers of linear equations are typically based on *multigrid* or *domain decomposition* methods. Both methods use coarse spaces to generate related well-conditioned systems, which can be solved efficiently by standard iterative solvers with the rate of convergence in terms of the condition number. The methods differ in the way the coarse grids are generated. The auxiliary coarse discretizations used by the multigrid are typically generated by the hierarchy of coarse discretisations of displacements, while those used by the domain decomposition are generated by the decomposition of the domain as in Fig. 1.3.

Here, we present theoretically supported scalable algorithms for contact problems that are based on variants of the FETI domain decomposition method. The scalability of the algorithms presented in this book is demonstrated by the solution of contact problems with billions of nodal variables. In the last chapter, it is indicated how the algorithms should be modified to be able to solve effectively contact problems discretized by hundreds of billions of nodal variables.

References

1. Hertz, H.R.: On the contact of elastic solids. Journal für die reine und angewandte Mathematik. **92**, 156–171 (1981): In English translation: Hertz, H.: Miscellaneous Papers, pp. 146–183. Macmillan, New York (1896)
2. Fichera, G.: Problemi elastostatici con vincoli unilaterali II. Problema di Signorini con ambique condizioni al contorno. Memorie della Accademia Nazionale dei Lincei, S. VIII, vol. VII, Sez, I, 5 91–140 (1964)
3. Hlaváček, I., Haslinger, J., Nečas, J., Lovíšek, J.: Solution of Variational Inequalities in Mechanics. Springer, Berlin (1988)
4. Kikuchi, N., Oden, J.T.: Contact Problems in Elasticity. SIAM, Philadelphia (1988)
5. Coulomb, C.A.: Théorie des machines simples, en ayant égard au frottement de leurs parties et la raideur des cordages. Mémoirs Savants Étrangers **X**, 163–332 (1785)
6. Eck, C., Jarůšek, J., Krbec, M.: Unilateral Contact Problems. Chapman & Hall/CRC, London (2005)
7. Wohlmuth, B.I.: Variationally consistent discretization scheme and numerical algorithms for contact problems. Acta Numer. **20**, 569–734 (2011)

Part I
Basic Concepts

Chapter 2
Linear Algebra

The purpose of this chapter is to briefly review the notations, definitions, and results of linear algebra that are used in the rest of the book. There is no claim of completeness as the reader is assumed to be familiar with the basic concepts of college linear algebra such as vector spaces, linear mappings, matrix decompositions, etc. More systematic exposition and additional material can be found in the books by Demmel [1], Laub [2], or Golub and Van Loan [3].

2.1 Vectors and Matrices

In this book, we work with *n-dimensional arithmetic vectors* $\mathbf{v} \in \mathbb{R}^n$, where $\mathbb{R}$ denotes the set of real numbers. The only exception is Sect. 2.7, where vectors with complex entries are considered. We denote the ith component of an arithmetic vector $\mathbf{v} \in \mathbb{R}^n$ by $[\mathbf{v}]_i$. Thus $[\mathbf{v}]_i = v_i$ if $\mathbf{v} = [v_i]$ is defined by its components v_i. All the arithmetic vectors are considered by default to be column vectors. The relations between vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ are defined componentwise. Thus $\mathbf{u} \leq \mathbf{v}$ is equivalent to $[\mathbf{u}]_i \leq [\mathbf{v}]_i$, $i = 1, \ldots, n$. We sometimes call the elements of $\mathbb{R}^n$ *points* to indicate that the concepts of length and direction are not important.

The vector analog of $0 \in \mathbb{R}$ is the *zero vector* $\mathbf{o}_n \in \mathbb{R}^n$ with all the entries equal to zero. When the dimension can be deduced from the context, possibly using the assumption that all the expressions in our book are well defined, we often drop the subscript and write simply $\mathbf{o}$.

Given vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^n$, the set

$$\mathrm{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\} = \{\mathbf{v} \in \mathbb{R}^n : \mathbf{v} = \alpha_1 \mathbf{v}_1 + \cdots + \alpha_k \mathbf{v}_k, \ \alpha_i \in \mathbb{R}\}$$

is a vector space called the *linear span* of $\mathbf{v}_1, \ldots, \mathbf{v}_k$. For example

$$\mathrm{Span}\{\mathbf{s}_1, \ldots, \mathbf{s}_n\} = \mathbb{R}^n, \quad [\mathbf{s}_i]_j = \delta_{ij}, \quad i, j = 1, \ldots, n,$$

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3_2

where δ_{ij} denotes the *Kronecker symbol* defined by $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ for $i \neq j$, is the *standard basis* of $\mathbb{R}^n$.

We sometimes use the componentwise extensions of scalar functions to vectors. Thus, if $\mathbf{v} \in \mathbb{R}^n$, then $\mathbf{v}^+$ and $\mathbf{v}^-$ are the vectors the ith components of which are $\max\{[\mathbf{v}]_i, 0\}$ and $\min\{[\mathbf{v}]_i, 0\}$, respectively.

If $\mathcal{I}$ is a nonempty subset of $\{1, \ldots, n\}$ and $\mathbf{v} \in \mathbb{R}^n$, then we denote by $[\mathbf{v}]_\mathcal{I}$ or simply $\mathbf{v}_\mathcal{I}$ the subvector of $\mathbf{v}$ with components $[\mathbf{v}]_i$, $i \in \mathcal{I}$. Thus if $\mathcal{I}$ has m elements, then $\mathbf{v}_\mathcal{I} \in \mathbb{R}^m$, so we can refer to the components of $\mathbf{v}_\mathcal{I}$ either by the *global indices* $i \in \mathcal{I}$ or by the *local indices* $j \in \{1, \ldots, m\}$. We usually rely on the reader's judgment to recognize the appropriate type of indexing.

Similarly to the related convention for vectors, the (i, j)th component of a matrix $\mathsf{A} \in \mathbb{R}^{m \times n}$ is denoted by $[\mathsf{A}]_{ij}$, so that $[\mathsf{A}]_{ij} = a_{ij}$ for $\mathsf{A} = [a_{ij}]$ which is defined by its entries a_{ij}. A matrix $\mathsf{A} \in \mathbb{R}^{m \times n}$ is called an (m, n)-*matrix*.

The matrix analog of 0 is the *zero matrix* $\mathsf{O}_{mn} \in \mathbb{R}^{m \times n}$ with all the entries equal to zero. When the dimension is clear from the context, we often drop the subscripts and write simply O.

The matrix counterpart of $1 \in \mathbb{R}$ in $\mathbb{R}^{n \times n}$ is the *identity matrix* $\mathsf{I}_n = [\delta_{ij}]$ of the order n. When the dimension may be deduced from the context, we often drop the subscripts and write simply I. Thus, we can write

$$\mathsf{A} = \mathsf{I}\mathsf{A} = \mathsf{A}\mathsf{I}$$

for any matrix A, having in mind that the order of I on the left may be different from that on the right.

A matrix A is *positive definite* if $\mathbf{x}^T \mathsf{A}\mathbf{x} > 0$ for any $\mathbf{x} \neq \mathbf{o}$, *positive semidefinite* if $\mathbf{x}^T \mathsf{A}\mathbf{x} \geq 0$ for any $\mathbf{x}$, and *indefinite* if neither A nor $-\mathsf{A}$ is positive definite or semidefinite. We are especially interested in *symmetric positive definite (SPD)* or *symmetric positive semidefinite (SPS)* matrices.

If $\mathsf{A} \in \mathbb{R}^{m \times n}$, $\mathcal{I} \subseteq \{1, \ldots, m\}$, and $\mathcal{J} \subseteq \{1, \ldots, n\}$, $\mathcal{I}$ and $\mathcal{J}$ nonempty, we denote by $\mathsf{A}_{\mathcal{I}\mathcal{J}}$ the submatrix of A with the components $[\mathsf{A}]_{ij}$, $i \in \mathcal{I}$, $j \in \mathcal{J}$. The local indexing of the entries of $\mathsf{A}_{\mathcal{I}\mathcal{J}}$ is used whenever it is convenient in a similar way as the local indexing of subvectors which was introduced in Sect. 2.1. The full set of indices may be replaced by * so that $\mathsf{A} = \mathsf{A}_{**}$ and $\mathsf{A}_{\mathcal{I}*}$ denotes the submatrix of A with the row indices belonging to $\mathcal{I}$. Occasionally we simplify $\mathsf{A}_\mathcal{I} = \mathsf{A}_{\mathcal{I}*}$.

Sometimes it is useful to rearrange the matrix operations into manipulations with submatrices of given matrices called blocks. A *block matrix* $\mathsf{A} \in \mathbb{R}^{m \times n}$ is defined by its blocks $\mathsf{A}_{ij} = \mathsf{A}_{\mathcal{I}_i \mathcal{J}_j}$, where $\mathcal{I}_i$ and $\mathcal{J}_j$ denote nonempty contiguous sets of indices decomposing $\{1, \ldots, m\}$ and $\{1, \ldots, n\}$, respectively. We can use the block structure to implement matrix operations only when the block structure of the involved matrices matches.

The matrices in our applications are often *sparse* in the sense that they have a small number of nonzero entries distributed in a pattern which can be exploited to the efficient implementation of matrix operations or to the reduction of storage requirements.

2.2 Matrices and Mappings

Each matrix $\mathsf{A} \in \mathbb{R}^{m\times n}$ defines the mapping which assigns to each $\mathbf{x} \in \mathbb{R}^n$ the vector $\mathsf{A}\mathbf{x} \in \mathbb{R}^m$. Two important subspaces associated with this mapping are its *range* or *image space* $\mathrm{Im}\mathsf{A}$ and its *kernel* or *null space* $\mathrm{Ker}\mathsf{A}$; they are defined by

$$\mathrm{Im}\mathsf{A} = \{\mathsf{A}\mathbf{x} : \ \mathbf{x} \in \mathbb{R}^n\} \quad \text{and} \quad \mathrm{Ker}\mathsf{A} = \{\mathbf{x} \in \mathbb{R}^n : \ \mathsf{A}\mathbf{x} = \mathbf{o}\}.$$

The range of A is the span of its columns. The *rank* and the *defect* of a matrix are defined as the dimension of its image and kernel, respectively.

If f is a mapping defined on $\mathscr{D} \subseteq \mathbb{R}^n$ and $\Omega \subseteq \mathscr{D}$, then $f|\Omega$ denotes the *restriction* of f to Ω, that is, the mapping defined on Ω which assigns to each $\mathbf{x} \in \Omega$ the value $f(\mathbf{x})$. If $\mathsf{A} \in \mathbb{R}^{m\times n}$ and V is a subspace of $\mathbb{R}^n$, we define $\mathsf{A}|V$ as a restriction of the mapping associated with A to V. The restriction $\mathsf{A}|V$ is said to be positive definite if $\mathbf{x}^T\mathsf{A}\mathbf{x} > 0$ for $\mathbf{x} \in V$, $\mathbf{x} \neq \mathbf{o}$, and positive semidefinite if $\mathbf{x}^T\mathsf{A}\mathbf{x} \geq 0$ for $\mathbf{x} \in V$.

The mapping associated with A is *injective* if $\mathsf{A}\mathbf{x} = \mathsf{A}\mathbf{y}$ implies $\mathbf{x} = \mathbf{y}$. It is easy to check that the mapping associated with A is injective if and only if $\mathrm{Ker}\mathsf{A} = \{\mathbf{o}\}$. If $m = n$, then A is injective if and only if $\mathrm{Im}\mathsf{A} = \mathbb{R}^n$.

A subspace $V \subseteq \mathbb{R}^n$ which satisfies

$$\mathsf{A}V = \{\mathsf{A}\mathbf{x} : \ \mathbf{x} \in V\} \subseteq V$$

is an *invariant subspace* of A. Obviously

$$\mathsf{A}(\mathrm{Im}\mathsf{A}) \subseteq \mathrm{Im}\mathsf{A},$$

so that $\mathrm{Im}\mathsf{A}$ is an invariant subspace of A.

A *projector* is a square matrix P that satisfies

$$\mathsf{P}^2 = \mathsf{P}.$$

A vector $\mathbf{x} \in \mathrm{Im}\mathsf{P}$ if and only if there is $\mathbf{y} \in \mathbb{R}^n$ such that $\mathbf{x} = \mathsf{P}\mathbf{y}$, so that

$$\mathsf{P}\mathbf{x} = \mathsf{P}(\mathsf{P}\mathbf{y}) = \mathsf{P}\mathbf{y} = \mathbf{x}.$$

If P is a projector, then also $\mathsf{Q} = \mathsf{I} - \mathsf{P}$ and P^T are projectors as

$$(\mathsf{I} - \mathsf{P})^2 = \mathsf{I} - 2\mathsf{P} + \mathsf{P}^2 = \mathsf{I} - \mathsf{P} \quad \text{and} \quad \left(\mathsf{P}^T\right)^2 = \left(\mathsf{P}^2\right)^T = \mathsf{P}^T.$$

Since for any $\mathbf{x} \in \mathbb{R}^n$

$$\mathbf{x} = \mathsf{P}\mathbf{x} + (\mathsf{I} - \mathsf{P})\mathbf{x},$$

it simply follows that $\mathrm{Im}\mathsf{Q} = \mathrm{Ker}\mathsf{P}$,

$$\mathbb{R}^n = \mathrm{Im}\mathsf{P} + \mathrm{Ker}\mathsf{P}, \quad \text{and} \quad \mathrm{Ker}\mathsf{P} \cap \mathrm{Im}\mathsf{P} = \{\mathbf{o}\}.$$

We say that P is a projector onto $U = \mathrm{Im}\mathsf{P}$ along $V = \mathrm{Ker}\mathsf{P}$ and Q is a complementary projector onto V along U. The above relations may be rewritten as

$$\mathrm{Im}\mathsf{P} \oplus \mathrm{Ker}\mathsf{P} = \mathbb{R}^n. \tag{2.1}$$

Let $(\pi(1), \ldots, \pi(n))$ be a permutation of numbers $1, \ldots, n$. Then, the mapping which assigns to each $\mathbf{v} = [v_i] \in \mathbb{R}^n$ a vector $[v_{\pi(1)}, \ldots, v_{\pi(n)}]^T$ is associated with the *permutation matrix*

$$\mathsf{P} = [\mathbf{s}_{\pi(1)}, \ldots, \mathbf{s}_{\pi(n)}],$$

where $\mathbf{s}_i$ denotes the ith column of the identity matrix I_n. If P is a permutation matrix, then

$$\mathsf{P}\mathsf{P}^T = \mathsf{P}^T\mathsf{P} = \mathsf{I}.$$

Notice that if B is a matrix obtained from a matrix A by reordering of the rows of A, then there is a permutation matrix P such that $\mathsf{B} = \mathsf{P}\mathsf{A}$. Similarly, if B is a matrix obtained from A by reordering of the columns of A, then there is a permutation matrix P such that $\mathsf{B} = \mathsf{A}\mathsf{P}$.

2.3 Inverse and Generalized Inverse

If A is a square full rank matrix, then there is the unique *inverse matrix* A^{-1} such that

$$\mathsf{A}\mathsf{A}^{-1} = \mathsf{A}^{-1}\mathsf{A} = \mathsf{I}. \tag{2.2}$$

The mapping associated with A^{-1} is inverse to that associated with A.

If A^{-1} exists, we say that A is *nonsingular*. A square matrix is *singular* if its inverse matrix does not exist. If P is a permutation matrix, then P is nonsingular and

$$\mathsf{P}^{-1} = \mathsf{P}^T.$$

If A is a nonsingular matrix, then $\mathsf{A}^{-1}\mathbf{b}$ is the unique solution of $\mathsf{A}\mathbf{x} = \mathbf{b}$.

If A is nonsingular, then we can transpose (2.2) to get

$$(\mathsf{A}^{-1})^T\mathsf{A}^T = \mathsf{A}^T(\mathsf{A}^{-1})^T = \mathsf{I},$$

so that

$$(\mathsf{A}^T)^{-1} = (\mathsf{A}^{-1})^T. \tag{2.3}$$

It follows that if A is symmetric, then A^{-1} is symmetric.

If $\mathsf{A} \in \mathbb{R}^{n\times n}$ is positive definite, then also A^{-1} is positive definite, as any vector $\mathbf{x} \neq \mathbf{o}$ can be expressed as $\mathbf{x} = \mathsf{A}\mathbf{y}$, $\mathbf{y} \neq \mathbf{o}$, and

$$\mathbf{x}^T\mathsf{A}^{-1}\mathbf{x} = (\mathsf{A}\mathbf{y})^T\mathsf{A}^{-1}\mathsf{A}\mathbf{y} = \mathbf{y}^T\mathsf{A}^T\mathbf{y} = \mathbf{y}^T\mathsf{A}\mathbf{y} > 0.$$

If A and B are nonsingular matrices, then it is easy to check that also AB is nonsingular and

$$(\mathsf{AB})^{-1} = \mathsf{B}^{-1}\mathsf{A}^{-1}.$$

If

$$\mathsf{H} = \begin{bmatrix} \mathsf{H}_{\mathscr{I}\mathscr{I}} & \mathsf{H}_{\mathscr{I}\mathscr{J}} \\ \mathsf{H}_{\mathscr{J}\mathscr{I}} & \mathsf{H}_{\mathscr{J}\mathscr{J}} \end{bmatrix} = \begin{bmatrix} \mathsf{A} & \mathsf{B}^T \\ \mathsf{B} & \mathsf{C} \end{bmatrix}$$

is an SPD block matrix, then we can directly evaluate

$$\mathsf{H}^{-1} = \begin{bmatrix} \mathsf{A} & \mathsf{B}^T \\ \mathsf{B} & \mathsf{C} \end{bmatrix}^{-1} = \begin{bmatrix} \mathsf{A}^{-1} - \mathsf{A}^{-1}\mathsf{B}^T\mathsf{S}^{-1}\mathsf{B}\mathsf{A}^{-1}, & -\mathsf{A}^{-1}\mathsf{B}^T\mathsf{S}^{-1} \\ -\mathsf{S}^{-1}\mathsf{B}\mathsf{A}^{-1}, & \mathsf{S}^{-1} \end{bmatrix}, \tag{2.4}$$

where $\mathsf{S} = \mathsf{C} - \mathsf{B}\mathsf{A}^{-1}\mathsf{B}^T$ denotes the *Schur complement* of H with respect to A. Thus

$$\left[\mathsf{A}^{-1}\right]_{\mathscr{J}\mathscr{J}} = \mathsf{S}^{-1}. \tag{2.5}$$

If $\mathsf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathrm{Im}\mathsf{A}$, then we can express a solution of the system of linear equations $\mathsf{A}\mathbf{x} = \mathbf{b}$ by means of a *left generalized inverse matrix* $\mathsf{A}^+ \in \mathbb{R}^{n\times m}$ which satisfies $\mathsf{A}\mathsf{A}^+\mathsf{A} = \mathsf{A}$. Indeed, if $\mathbf{b} \in \mathrm{Im}\mathsf{A}$, then there is $\mathbf{y}$ such that $\mathbf{b} = \mathsf{A}\mathbf{y}$ and $\overline{\mathbf{x}} = \mathsf{A}^+\mathbf{b}$ satisfies

$$\mathsf{A}\overline{\mathbf{x}} = \mathsf{A}\mathsf{A}^+\mathbf{b} = \mathsf{A}\mathsf{A}^+\mathsf{A}\mathbf{y} = \mathsf{A}\mathbf{y} = \mathbf{b}.$$

Thus A^+ acts on the range of A like the inverse matrix. If A is a nonsingular square matrix, then obviously

$$\mathsf{A}^+ = \mathsf{A}^{-1}.$$

Moreover, if $\mathsf{A} \in \mathbb{R}^{n\times n}$ and $\mathsf{S} \in \mathbb{R}^{n\times p}$ are such that $\mathsf{A}\mathsf{S} = \mathsf{O}$, then $(\mathsf{A}^+) + \mathsf{S}\mathsf{S}^T$ is also a left generalized inverse as

$$\mathsf{A}\left(\mathsf{A}^+ + \mathsf{S}\mathsf{S}^T\right)\mathsf{A} = \mathsf{A}\mathsf{A}^+\mathsf{A} + \mathsf{A}\mathsf{S}\mathsf{S}^T\mathsf{A} = \mathsf{A}.$$

If A is a symmetric singular matrix, then there is a permutation matrix P such that

$$\mathsf{A} = \mathsf{P}^T\begin{bmatrix} \mathsf{B} & \mathsf{C}^T \\ \mathsf{C} & \mathsf{C}\mathsf{B}^{-1}\mathsf{C}^T \end{bmatrix}\mathsf{P},$$

where B is a nonsingular matrix the dimension of which is equal to the rank of A. It may be verified directly that the matrix

$$\mathsf{A}^{\#} = \mathsf{P}^T\begin{bmatrix} \mathsf{B}^{-1} & \mathsf{O}^T \\ \mathsf{O} & \mathsf{O} \end{bmatrix}\mathsf{P} \tag{2.6}$$

is a left generalized inverse of A. If A is SPS, then $\mathsf{A}^{\#}$ is also SPS. Notice that if $\mathsf{A}\mathsf{S} = \mathsf{O}$, then $\mathsf{A}^+ = \mathsf{A}^{\#} + \mathsf{S}\mathsf{S}^T$ is also an SPS generalized inverse.

2.4 Direct Methods for Solving Linear Equations

The inverse matrix is a useful tool for theoretical developments, but not for computations. It is often much more efficient to implement the multiplication of a vector by the inverse matrix by solving the related system of linear equations. We recall here briefly the *direct methods*, which reduce solving of the original system of linear equations to solving of a system or systems of equations with triangular matrices.

A matrix $\mathsf{L} = [l_{ij}]$ is *lower triangular* if $l_{ij} = 0$ for $i < j$. It is easy to solve a system $\mathsf{L}\mathbf{x} = \mathbf{b}$ with the nonsingular lower triangular matrix $\mathsf{L} \in \mathbb{R}^n$. As there is only one unknown in the first equation, we can find it and then substitute it into the remaining equations to obtain a system with the same structure, but with only $n - 1$ remaining unknowns. Repeating the procedure, we can find all the components of $\mathbf{x}$.

A similar procedure, but starting from the last equation, can be applied to a system with the nonsingular *upper triangular matrix* $\mathsf{U} = [u_{ij}]$ with $u_{ij} = 0$ for $i > j$.

The solution costs of a system with triangular matrices is proportional to the number of its nonzero entries. In particular, the solution of a system of linear equations with a *diagonal matrix* $\mathsf{D} = [d_{ij}]$, $d_{ij} = 0$ for $i \neq j$, reduces to the solution of a sequence of linear equations with one unknown.

If we are to solve the system of linear equations with a nonsingular matrix, we can use systematically *equivalent transformations* that do not change the solution in order to modify the original system to that with an upper triangular matrix. It is well known that the solutions of a system of linear equations are the same as the solutions of a system of linear equations obtained from the original system by interchanging two equations, replacing an equation by its nonzero multiple, or adding a multiple of one equation to another equation. The *Gauss elimination* for the solution of a system of linear equations with a nonsingular matrix thus consists of two steps: the *forward reduction*, which exploits equivalent transformations to reduce the original system to the system with an upper triangular matrix, and the *backward substitution*, which solves the resulting system with the upper triangular matrix.

Alternatively, we can use suitable matrix factorizations. For example, it is well known that any SPD matrix A can be decomposed into the product

$$\mathsf{A} = \mathsf{L}\mathsf{L}^T, \tag{2.7}$$

where L is a nonsingular lower triangular matrix with positive diagonal entries. Having the decomposition, we can evaluate $\mathbf{z} = \mathsf{A}^{-1}\mathbf{x}$ by solving the systems

$$\mathsf{L}\mathbf{y} = \mathbf{x} \quad \text{and} \quad \mathsf{L}^T\mathbf{z} = \mathbf{y}.$$

The factorization-based solvers may be especially useful when we are to solve several systems of equations with the same coefficients but different right-hand sides.

The method of evaluation of the factor L is known as the *Cholesky factorization*. The Cholesky factor L can be computed column by column. Suppose that

$$\mathsf{A} = \begin{bmatrix} a_{11} & \mathbf{a}_1^T \\ \mathbf{a}_1 & \mathsf{A}_{22} \end{bmatrix} \quad \text{and} \quad \mathsf{L} = \begin{bmatrix} l_{11} & \mathbf{o} \\ \mathbf{l}_1 & \mathsf{L}_{22} \end{bmatrix}.$$

Substituting for A and L into (2.7) and comparing the corresponding terms immediately reveals that

$$l_{11} = \sqrt{a_{11}}, \qquad \mathbf{l}_1 = l_{11}^{-1}\mathbf{a}_1, \qquad \mathsf{L}_{22}\mathsf{L}_{22}^T = \mathsf{A}_{22} - \mathbf{l}_1\mathbf{l}_1^T. \tag{2.8}$$

This gives us the first column of L, and the remaining factor L_{22} is simply the Cholesky factor of the Schur complement $\mathsf{A}_{22} - \mathbf{l}_1\mathbf{l}_1^T$ which is known to be positive definite, so we can find its first column by the above procedure. The algorithm can be implemented to exploit a sparsity pattern of A, e.g., when $\mathsf{A} = [a_{ij}] \in \mathbb{R}^{n\times n}$ is a *band matrix* with $a_{ij} = 0$ for $|i - j| > b$, $b \ll n$.

If $\mathsf{A} \in \mathbb{R}^{n\times n}$ is only positive semidefinite, it can happen that $a_{11} = 0$. Then

$$0 \le \mathbf{x}^T\mathsf{A}\mathbf{x} = \mathbf{y}^T\mathsf{A}_{22}\mathbf{y} + 2x_1\mathbf{a}_1^T\mathbf{y}$$

for any vector $\mathbf{x} = \left[x_1, \mathbf{y}^T\right]^T$. The inequality implies that $\mathbf{a}_1 = \mathbf{o}$, as otherwise we could take $\mathbf{y} = -\mathbf{a}_1$ and large x_1 to get

$$\mathbf{y}^T\mathsf{A}_{22}\mathbf{y} + 2x_1\mathbf{a}_1^T\mathbf{y} = \mathbf{a}_1^T\mathsf{A}_{22}\mathbf{a}_1 - 2x_1\|\mathbf{a}_1\|^2 < 0.$$

Thus for A symmetric positive semidefinite and $a_{11} = 0$, (2.8) reduces to

$$l_{11} = 0, \qquad \mathbf{l}_1 = \mathbf{o}, \qquad \mathsf{L}_{22}\mathsf{L}_{22}^T = \mathsf{A}_{22}. \tag{2.9}$$

This simple modification assumes exact arithmetics. In the computer arithmetics, the decision whether a_{11} is to be treated as zero depends on some small $\varepsilon > 0$. Alternatively, it is possible to exploit some additional information. For example, any orthonormal basis of the kernel of a matrix can be used to identify the zero rows (and columns) of a Cholesky factor by means of the following lemma.

Lemma 2.1 *Let $\mathsf{A} \in \mathbb{R}^{n\times n}$ denote an SPS matrix the kernel of which is spanned by the full column rank matrix $\mathsf{R} \in \mathbb{R}^{n\times d}$ with orthonormal columns. Let*

$$\mathscr{I} = \{i_1, \ldots, i_d\}, \quad 1 \le i_1 < i_2 < \cdots < i_d \le n,$$

denote a set of indices, and let $\mathscr{J} = \mathscr{N} - \mathscr{I}$, $\mathscr{N} = \{1, 2, \ldots, n\}$.

Then

$$\lambda_{\min}(\mathsf{A}_{\mathscr{J}\mathscr{J}}) \ge \overline{\lambda}_{\min}(\mathsf{A})\,\sigma_{\min}^4(\mathsf{R}_{\mathscr{J}*}), \tag{2.10}$$

where $\overline{\lambda}_{\min}(\mathsf{A})$ and $\sigma_{\min}(\mathsf{R}_{\mathscr{J}*})$ denote the least nonzero eigenvalue of A and the least singular value of $\mathsf{R}_{\mathscr{J}*}$, respectively.

Proof See Dostál et al. [4]. □

If we can identify a nonsingular block $\mathsf{A}_{\mathcal{J}\mathcal{J}}$ of A, then we can reduce the problems related to the manipulation with singular matrices to those with smaller matrices using the decomposition

$$\mathsf{A} = \begin{bmatrix} \mathsf{A}_{\mathcal{J}\mathcal{J}} & \mathsf{A}_{\mathcal{J}\mathcal{I}} \\ \mathsf{A}_{\mathcal{I}\mathcal{J}} & \mathsf{A}_{\mathcal{I}\mathcal{I}} \end{bmatrix} = \begin{bmatrix} \mathsf{L}_{\mathcal{J}\mathcal{J}} & \mathsf{O} \\ \mathsf{L}_{\mathcal{I}\mathcal{J}} & \mathsf{I} \end{bmatrix} \begin{bmatrix} \mathsf{L}^T_{\mathcal{J}\mathcal{J}} & \mathsf{L}^T_{\mathcal{I}\mathcal{J}} \\ \mathsf{O} & \mathsf{S} \end{bmatrix}, \tag{2.11}$$

where $\mathsf{L}_{\mathcal{J}\mathcal{J}} \in \mathbb{R}^{r\times r}$ is a lower factor of the Cholesky decomposition of $\mathsf{A}_{\mathcal{J}\mathcal{J}}$, $\mathsf{L}_{\mathcal{I}\mathcal{J}} \in \mathbb{R}^{s\times r}$, $r = n - s$, $\mathsf{L}_{\mathcal{I}\mathcal{J}} = \mathsf{A}_{\mathcal{I}\mathcal{J}}\mathsf{L}^{-T}_{\mathcal{J}\mathcal{J}}$, and $\mathsf{S} \in \mathbb{R}^{s\times s}$ is the *Schur complement matrix of the block* $\mathsf{A}_{\mathcal{I}\mathcal{I}}$ defined by

$$\mathsf{S} = \mathsf{A}_{\mathcal{I}\mathcal{I}} - \mathsf{A}_{\mathcal{I}\mathcal{J}}\mathsf{A}^{-1}_{\mathcal{J}\mathcal{J}}\mathsf{A}_{\mathcal{J}\mathcal{I}}.$$

The decomposition (2.11) is a useful tool for the effective construction of a generalized inverse or for the effective evaluation of the multiplication of a vector by a generalized inverse.

Lemma 2.2 *Let* A *denote an SPS block matrix as in (2.11), let* $\mathbf{e} \in \mathrm{Ker}\mathsf{A}$*, and let* S *denote the Schur complement matrix of the block* $\mathsf{A}_{\mathcal{I}\mathcal{I}}$*. Then* $\mathbf{e}_{\mathcal{J}} \in \mathrm{Ker}\mathsf{S}$ *and if* S^+ *denote any generalized inverse of* S*, then the matrix* A^+ *defined by*

$$\mathsf{A}^+ = \begin{bmatrix} \mathsf{L}^{-T}_{\mathcal{J}\mathcal{J}} & -\mathsf{L}^{-T}_{\mathcal{J}\mathcal{J}}\mathsf{L}^T_{\mathcal{I}\mathcal{J}}\mathsf{S}^+ \\ \mathsf{O} & \mathsf{S}^+ \end{bmatrix} \begin{bmatrix} \mathsf{L}^{-1}_{\mathcal{J}\mathcal{J}} & \mathsf{O} \\ -\mathsf{L}_{\mathcal{I}\mathcal{J}}\mathsf{L}^{-1}_{\mathcal{J}\mathcal{J}} & \mathsf{I} \end{bmatrix} \tag{2.12}$$

is a generalized inverse of A *which satisfies*

$$[\mathsf{A}^+]_{\mathcal{I}\mathcal{I}} = \mathsf{S}^+. \tag{2.13}$$

Proof If $\mathsf{A}\mathbf{e} = \mathbf{o}$, then

$$\mathsf{A}_{\mathcal{J}\mathcal{J}}\mathbf{e}_{\mathcal{J}} + \mathsf{A}_{\mathcal{J}\mathcal{I}}\mathbf{e}_{\mathcal{I}} = \mathbf{o}, \qquad \mathsf{A}_{\mathcal{I}\mathcal{J}}\mathbf{e}_{\mathcal{J}} + \mathsf{A}_{\mathcal{I}\mathcal{I}}\mathbf{e}_{\mathcal{I}} = \mathbf{o},$$

and

$$\begin{aligned} \mathsf{S}\mathbf{e}_{\mathcal{I}} &= (\mathsf{A}_{\mathcal{I}\mathcal{I}} - \mathsf{A}_{\mathcal{I}\mathcal{J}}\mathsf{A}^{-1}_{\mathcal{J}\mathcal{J}}\mathsf{A}_{\mathcal{J}\mathcal{I}})\mathbf{e}_{\mathcal{I}} \\ &= \mathsf{A}_{\mathcal{I}\mathcal{I}}\mathbf{e}_{\mathcal{I}} - \mathsf{A}_{\mathcal{I}\mathcal{J}}\mathsf{A}^{-1}_{\mathcal{J}\mathcal{J}}(-\mathsf{A}_{\mathcal{J}\mathcal{J}}\mathbf{e}_{\mathcal{J}}) = \mathbf{o}, \end{aligned}$$

i.e., $\mathbf{e}_{\mathcal{J}} \in \mathrm{Ker}\mathsf{S}$. The rest can be verified directly. □

Notice that

$$\mathsf{S}^{+} = \left[\mathsf{A}^{+}\right]_{\mathscr{J}\,\mathscr{J}}. \tag{2.14}$$

If A is a full rank matrix, then we get

$$\mathsf{S}^{-1} = \left[\mathsf{A}^{-1}\right]_{\mathscr{J}\,\mathscr{J}}, \tag{2.15}$$

which agrees with (2.5).

2.5 Norms

General concepts of size and distance in a vector space are expressed by norms. A *norm* on $\mathbb{R}^n$ is a function which assigns to each $\mathbf{x} \in \mathbb{R}^n$ a number $\mathbb{I}\mathbf{x}\mathbb{I} \in \mathbb{R}$ in such a way that for any vectors $\mathbf{x},\ \mathbf{y} \in \mathbb{R}^n$ and any scalar $\alpha \in \mathbb{R}$, the following three conditions are satisfied:

(i) $\mathbb{I}\mathbf{x}\mathbb{I} \geq 0$, and $\mathbb{I}\mathbf{x}\mathbb{I} = 0$ if and only if $\mathbf{x} = \mathbf{o}$.
(ii) $\mathbb{I}\mathbf{x} + \mathbf{y}\mathbb{I} \leq \mathbb{I}\mathbf{x}\mathbb{I} + \mathbb{I}\mathbf{y}\mathbb{I}$.
(iii) $\mathbb{I}\alpha\mathbf{x}\mathbb{I} = |\alpha|\ \mathbb{I}\mathbf{x}\mathbb{I}$.

It is easy to check that the functions

$$\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}_1^2 + \cdots + \mathbf{x}_n^2} \quad \text{and} \quad \|\mathbf{x}\|_\infty = \max\{|\mathbf{x}_1|, \ldots, |\mathbf{x}_n|\}$$

are norms. They are called ℓ_2 (Euclidean) and ℓ_∞ norms, respectively.

Given a norm defined on the domain and the range of a matrix A, we can define the *induced norm* $\mathbb{I}\mathsf{A}\mathbb{I}$ of A by

$$\mathbb{I}\mathsf{A}\mathbb{I} = \sup_{\mathbb{I}\mathbf{x}\mathbb{I}=1} \mathbb{I}\mathsf{A}\mathbf{x}\mathbb{I} = \sup_{\mathbf{x}\neq\mathbf{o}} \frac{\mathbb{I}\mathsf{A}\mathbf{x}\mathbb{I}}{\mathbb{I}\mathbf{x}\mathbb{I}}.$$

If $\mathsf{B} \neq \mathsf{O}$, then

$$\mathbb{I}\mathsf{A}\mathsf{B}\mathbb{I} = \sup_{\mathbf{x}\neq\mathbf{o}} \frac{\mathbb{I}\mathsf{A}\mathsf{B}\mathbf{x}\mathbb{I}}{\mathbb{I}\mathbf{x}\mathbb{I}} = \sup_{\mathsf{B}\mathbf{x}\neq\mathbf{o}} \frac{\mathbb{I}\mathsf{A}\mathsf{B}\mathbf{x}\mathbb{I}}{\mathbb{I}\mathsf{B}\mathbf{x}\mathbb{I}} \frac{\mathbb{I}\mathsf{B}\mathbf{x}\mathbb{I}}{\mathbb{I}\mathbf{x}\mathbb{I}} \leq \sup_{\substack{\mathbf{y}\in\mathrm{Im}\mathsf{B},\\ \mathbf{y}\neq\mathbf{o}}} \frac{\mathbb{I}\mathsf{A}\mathbf{y}\mathbb{I}}{\mathbb{I}\mathbf{y}\mathbb{I}} \sup_{\mathbf{x}\neq\mathbf{o}} \frac{\mathbb{I}\mathsf{B}\mathbf{x}\mathbb{I}}{\mathbb{I}\mathbf{x}\mathbb{I}}.$$

It follows easily that the induced norm is *submultiplicative*, i.e.,

$$\mathbb{I}\mathsf{A}\mathsf{B}\mathbb{I} \leq \mathbb{I}\mathsf{A}|\mathrm{Im}\mathsf{B}\mathbb{I}\ \mathbb{I}\mathsf{B}\mathbb{I} \leq \mathbb{I}\mathsf{A}\mathbb{I}\ \mathbb{I}\mathsf{B}\mathbb{I}. \tag{2.16}$$

If $\mathsf{A} = [a_{ij}] \in \mathbb{R}^{m\times n}$ and $\mathbf{x} = [x_i] \in \mathbb{R}^n$, then

$$\|\mathsf{A}\mathbf{x}\|_\infty = \max_{i=1,\dots,m} |\sum_{j=1}^{n} a_{ij}x_j| \le \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}||x_j| \le \|\mathbf{x}\|_\infty \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}|,$$

that is, $\|\mathsf{A}\|_\infty \le \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}|$. Since the last inequality turns into the equality for a vector $\mathbf{x}$ with suitably chosen entries $x_i \in \{1, -1\}$, we have

$$\|\mathsf{A}\|_\infty = \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}|. \tag{2.17}$$

2.6 Scalar Products

General concepts of length and angle in a vector space are introduced by means of a *scalar product*; it is the mapping which assigns to each couple $\mathbf{x},\ \mathbf{y} \in \mathbb{R}^n$ a number $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}$ in such a way that for any vectors $\mathbf{x},\ \mathbf{y},\ \mathbf{z} \in \mathbb{R}^n$ and any scalar $\alpha \in \mathbb{R}$, the following four conditions are satisfied:

(i) $(\mathbf{x}, \mathbf{y} + \mathbf{z}) = (\mathbf{x}, \mathbf{y}) + (\mathbf{x}, \mathbf{z})$.
(ii) $(\alpha\mathbf{x}, \mathbf{y}) = \alpha(\mathbf{x}, \mathbf{y})$.
(iii) $(\mathbf{x}, \mathbf{y}) = (\mathbf{y}, \mathbf{x})$.
(iv) $(\mathbf{x}, \mathbf{x}) > 0$ for $\mathbf{x} \neq \mathbf{o}$.

The scalar product is an SPD form, see also Chap. 4.

We often use the *Euclidean scalar product* or the *Euclidean inner product* which assigns to each couple of vectors $\mathbf{x},\ \mathbf{y} \in \mathbb{R}^n$ a number defined by

$$(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}.$$

In more complicated expressions, we often denote the Euclidean scalar product in $\mathbb{R}^3$ by dot, so that

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T\mathbf{y}.$$

If A is an SPD matrix, then we can define the more general A-*scalar product* on $\mathbb{R}^n$ by

$$(\mathbf{x}, \mathbf{y})_\mathsf{A} = \mathbf{x}^T\mathsf{A}\mathbf{y}.$$

We denote for any $\mathbf{x} \in \mathbb{R}^n$ its *Euclidean norm* and A-*norm* by

$$\|\mathbf{x}\| = (\mathbf{x}, \mathbf{x})^{1/2}, \quad \|\mathbf{x}\|_\mathsf{A} = (\mathbf{x}, \mathbf{x})_\mathsf{A}^{1/2}.$$

It is easy to see that any norm induced by a scalar product satisfies the properties (i) and (iii) of the norm. The property (ii) follows from the *Cauchy–Schwarz inequality*

$$(\mathbf{x}, \mathbf{y})^2 \leq \|\mathbf{x}\|^2 \, \|\mathbf{y}\|^2, \tag{2.18}$$

which is valid for any $\mathbf{x}, \ \mathbf{y} \in \mathbb{R}^n$ and any scalar product. The bound is tight in the sense that the inequality becomes the equality when $\mathbf{x}, \ \mathbf{y}$ are dependent.

A pair of vectors $\mathbf{x}$ and $\mathbf{y}$ is *orthogonal* (with respect to a given scalar product) if

$$(\mathbf{x}, \mathbf{y}) = 0.$$

The vectors $\mathbf{x}$ and $\mathbf{y}$ that are orthogonal in A-scalar product are called A-*conjugate* or briefly *conjugate*.

Two *sets of vectors* $\mathscr{E}$ *and* $\mathscr{F}$ *are orthogonal* (also stated "$\mathscr{E}$ orthogonal to $\mathscr{F}$") if any $\mathbf{x} \in \mathscr{E}$ is orthogonal to any $\mathbf{y} \in \mathscr{F}$. The set $\mathscr{E}^\perp$ of all the vectors of $\mathbb{R}^n$ that are orthogonal to $\mathscr{E} \subseteq \mathbb{R}^n$ is a vector space called an *orthogonal complement* of $\mathscr{E}$. If $\mathscr{E} \subseteq \mathbb{R}^n$, then

$$\mathbb{R}^n = \operatorname{Span} \mathscr{E} \oplus \mathscr{E}^\perp.$$

A *set of vectors* $\mathscr{E}$ *is orthogonal* if its elements are pairwise orthogonal, i.e., any $\mathbf{x} \in \mathscr{E}$ is orthogonal to any $\mathbf{y} \in \mathscr{E}$, $\mathbf{y} \neq \mathbf{x}$. A set of vectors $\mathscr{E}$ is *orthonormal* if it is orthogonal and $(\mathbf{x}, \mathbf{x}) = 1$ for any $\mathbf{x} \in \mathscr{E}$.

A square matrix U *is orthogonal* if $\mathsf{U}^T\mathsf{U} = \mathsf{I}$, that is, $\mathsf{U}^{-1} = \mathsf{U}^T$. Multiplication by an orthogonal matrix U preserves both the angles between any two vectors and the Euclidean norm of any vector as

$$(\mathsf{U}\mathbf{x})^T \mathsf{U}\mathbf{y} = \mathbf{x}^T \mathsf{U}^T \mathsf{U}\mathbf{y} = \mathbf{x}^T \mathbf{y}.$$

A matrix $\mathsf{P} \in \mathbb{R}^{n \times n}$ is an *orthogonal projector* if P is a projector, i.e., $\mathsf{P}^2 = \mathsf{P}$, and $\mathrm{Im}\mathsf{P}$ is orthogonal to $\mathrm{Ker}\mathsf{P}$. The latter condition can be rewritten equivalently as

$$\mathsf{P}^T(\mathsf{I} - \mathsf{P}) = \mathsf{O}.$$

It simply follows that

$$\mathsf{P}^T = \mathsf{P}^T\mathsf{P} = \mathsf{P},$$

so that orthogonal projectors are symmetric matrices and symmetric projectors are orthogonal projectors. If P is an orthogonal projector, then $\mathsf{I} - \mathsf{P}$ is also an orthogonal projector as

$$(\mathsf{I} - \mathsf{P})^2 = I - 2\mathsf{P} + \mathsf{P}^2 = \mathsf{I} - \mathsf{P} \quad \text{and} \quad (\mathsf{I} - \mathsf{P})^T\mathsf{P} = (\mathsf{I} - \mathsf{P})\mathsf{P} = \mathsf{O}.$$

If $U \subseteq \mathbb{R}^n$ is the subspace spanned by the columns of a full column rank matrix $\mathsf{U} \in \mathbb{R}^{m \times n}$, then

$$\mathsf{P} = \mathsf{U}(\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T$$

is an orthogonal projector as

$$\mathsf{P}^2 = \mathsf{U}(\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T\mathsf{U}(\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T = \mathsf{P} \quad \text{and} \quad \mathsf{P}^T = \mathsf{P}.$$

Since any vector $\mathbf{x} \in U$ may be written in the form $\mathbf{x} = \mathsf{U}\mathbf{y}$ and

$$\mathsf{P}\mathbf{x} = \mathsf{U}(\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T\mathsf{U}\mathbf{y} = \mathsf{U}\mathbf{y} = \mathbf{x},$$

it follows that

$$U = \mathrm{Im}\mathsf{P}.$$

Observe that $\mathsf{U}^T\mathsf{U}$ is nonsingular; since $\mathsf{U}^T\mathsf{U}\mathbf{x} = \mathbf{o}$ implies

$$\|\mathsf{U}\mathbf{x}\|^2 = \mathbf{x}^T(\mathsf{U}^T\mathsf{U}\mathbf{x}) = 0,$$

it follows that $\mathbf{x} = \mathbf{o}$ by the assumption on the full column rank of U.

2.7 Eigenvalues and Eigenvectors

Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ denote a square matrix with complex entries. If a vector $\mathbf{e} \in \mathbb{C}^n$ and a scalar $\lambda \in \mathbb{C}$ satisfy

$$\mathsf{A}\mathbf{e} = \lambda\mathbf{e}, \tag{2.19}$$

then $\mathbf{e}$ is said to be an *eigenvector* of A associated with an *eigenvalue* λ. A vector $\mathbf{e}$ is an eigenvector of A if and only if $\mathrm{Span}\{\mathbf{e}\}$ is an invariant subspace of A; the restriction $\mathsf{A}|\mathrm{Span}\{\mathbf{e}\}$ reduces to the multiplication by λ. If $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ are eigenvectors of a symmetric matrix A, then it is easy to check that $\mathrm{Span}\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ and $\mathrm{Span}\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}^\perp$ are invariant subspaces.

The set of all eigenvalues of A is called the *spectrum* of A; we denote it by $\sigma(\mathsf{A})$. Obviously, $\lambda \in \sigma(\mathsf{A})$ if and only if $\mathsf{A} - \lambda\mathsf{I}$ is singular, and $0 \in \sigma(\mathsf{A})$ if and only if A is singular. If $\lambda \neq 0$, $\lambda \in \sigma(\mathsf{A})$, then we can multiply (2.19) by $\lambda^{-1}\mathsf{A}^{-1}$ to get $\mathsf{A}^{-1}\mathbf{e} = \lambda^{-1}\mathbf{e}$, so we can write

$$\sigma(\mathsf{A}^{-1}) = \sigma^{-1}(\mathsf{A}).$$

If $U \subseteq \mathbb{C}^n$ is an invariant subspace of $\mathsf{A} \in \mathbb{C}^{n\times n}$, then we denote by $\sigma(\mathsf{A}|U)$ the eigenvalues of A that correspond to the eigenvectors belonging to U.

Since it is well known that a matrix is singular if and only if its determinant is equal to zero, it follows that the eigenvalues of A are the roots of the *characteristic equation*

$$\det(\mathsf{A} - \lambda\mathsf{I}) = 0. \tag{2.20}$$

The *characteristic polynomial* $p_{\mathsf{A}}(\lambda) = \det(\mathsf{A} - \lambda\mathsf{I})$ is of the degree n. Thus there are at most n distinct eigenvalues and $\sigma(\mathsf{A})$ is not the empty set.

Even though it is in general difficult to evaluate the eigenvalues of a given matrix A, it is still possible to get nontrivial information about $\sigma(\mathsf{A})$ without heavy computations. Useful information about the location of eigenvalues can be obtained by *Gershgorin's theorem*, which guarantees that every eigenvalue of $\mathsf{A} = [a_{ij}] \in \mathbb{C}^{n\times n}$ is located in at least one of the n circular disks in the complex plane with the centers a_{ii} and radii $r_i = \sum_{j\neq i} |a_{ij}|$.

The *eigenvalues of a real symmetric matrix are real*. Since it is easy to check whether a matrix is symmetric, this gives us useful information about the location of eigenvalues.

Let $\mathsf{A} \in \mathbb{R}^{n\times n}$ denote a real symmetric matrix, let $\mathscr{I} = \{1, \ldots, n-1\}$, and let $\mathsf{A}^1 = \mathsf{A}_{\mathscr{I}\mathscr{I}}$. Let $\lambda_1 \geq \cdots \geq \lambda_n$ and $\lambda_1^1 \geq \cdots \geq \lambda_{n-1}^1$ denote the eigenvalues of A and A^1, respectively. Then by the *Cauchy interlacing theorem*

$$\lambda_1 \geq \lambda_1^1 \geq \lambda_2 \geq \lambda_2^1 \geq \cdots \geq \lambda_{n-1}^1 \geq \lambda_n. \tag{2.21}$$

2.8 Matrix Decompositions

If $\mathsf{A} \in \mathbb{R}^{n\times n}$ is a symmetric matrix, then it is possible to find n orthonormal eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$ that form the basis of $\mathbb{R}^n$. Moreover, the corresponding eigenvalues are real. Denoting by $\mathsf{U} = [\mathbf{e}_1, \ldots, \mathbf{e}_n] \in \mathbb{R}^{n\times n}$ an orthogonal matrix the columns of which are the eigenvectors, we may write the *spectral decomposition* of A as

$$\mathsf{A} = \mathsf{U}\mathsf{D}\mathsf{U}^T, \tag{2.22}$$

where $\mathsf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n\times n}$ is the diagonal matrix the diagonal entries of which are the eigenvalues corresponding to the eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$. Reordering the columns of U, we can achieve that $\lambda_1 \geq \cdots \geq \lambda_n$.

The spectral decomposition reveals close relations between the properties of a symmetric matrix and its eigenvalues. Thus, a symmetric matrix is SPD if and only if all its eigenvalues are positive, and it is SPS if and only if they are nonnegative. The rank of a symmetric matrix is equal to the number of nonzero entries of D.

If A is symmetric, then we can use the spectral decomposition (2.22) to check that for any nonzero $\mathbf{x}$

$$\lambda_1 = \lambda_{\max} \geq \|\mathbf{x}\|^{-2}\mathbf{x}^T\mathsf{A}\mathbf{x} \geq \lambda_{\min} = \lambda_n. \tag{2.23}$$

Thus for any symmetric positive definite matrix A

$$\|\mathsf{A}\| = \lambda_{\max},\ \|\mathsf{A}^{-1}\| = \lambda_{\min}^{-1},\ \|\mathbf{x}\|_{\mathsf{A}} \leq \lambda_{\max}\|\mathbf{x}\|,\ \|\mathbf{x}\|_{\mathsf{A}^{-1}} \leq \lambda_{\min}^{-1}\|\mathbf{x}\|. \tag{2.24}$$

The *spectral condition number* $\kappa(\mathsf{A}) = \|\mathsf{A}\|\|\mathsf{A}^{-1}\|$, which is a measure of departure from the identity, can be expressed for real symmetric matrix by

$$\kappa(\mathsf{A}) = \lambda_{\max}/\lambda_{\min}.$$

If A is a real symmetric matrix and f is a real function defined on $\sigma(\mathsf{A})$, we can use the spectral decomposition to define the *scalar function* by

$$f(\mathsf{A}) = \mathsf{U} f(\mathsf{D})\mathsf{U}^T,$$

where $f(\mathsf{D}) = \operatorname{diag}(f(\lambda_1), \ldots, f(\lambda_n))$. It is easy to check that if a is the identity function on $\mathbb{R}$ defined by $a(x) = x$, then

$$a(\mathsf{A}) = \mathsf{A},$$

and if f and g are real functions defined on $\sigma(\mathsf{A})$, then

$$(f+g)(\mathsf{A}) = f(\mathsf{A}) + g(\mathsf{A}) \quad \text{and} \quad (f \cdot g)(\mathsf{A}) = f(\mathsf{A})g(\mathsf{A}).$$

Moreover, if $f(x) \geq 0$ for $x \in \sigma(\mathsf{A})$, then $f(\mathsf{A})$ is SPS, and if $f(x) > 0$ for $x \in \sigma(\mathsf{A})$, then $f(\mathsf{A})$ is SPD. For example, if A is SPD, then

$$\mathsf{A} = \mathsf{A}^{1/2}\mathsf{A}^{1/2}.$$

Obviously

$$\sigma(f(\mathsf{A})) = f(\sigma(\mathsf{A})), \tag{2.25}$$

and if $\mathbf{e}_i$ is an eigenvector corresponding to $\lambda_i \in \sigma(\mathsf{A})$, then it is also an eigenvector of $f(\mathsf{A})$ corresponding to $f(\lambda_i)$. It follows easily that for any SPS matrix

$$\operatorname{Im}\mathsf{A} = \operatorname{Im}\mathsf{A}^{1/2} \quad \text{and} \quad \operatorname{Ker}\mathsf{A} = \operatorname{Ker}\mathsf{A}^{1/2}. \tag{2.26}$$

A key to understanding nonsymmetric matrices is the *singular value decomposition* (SVD). If $\mathsf{B} \in \mathbb{R}^{m\times n}$, then SVD of B is given by

$$\mathsf{B} = \mathsf{U}\mathsf{S}\mathsf{V}^T, \tag{2.27}$$

where $\mathsf{U} \in \mathbb{R}^{m\times m}$ and $\mathsf{V} \in \mathbb{R}^{n\times n}$ are orthogonal, and $\mathsf{S} \in \mathbb{R}^{m\times n}$ is a diagonal matrix with nonnegative diagonal entries $\sigma_1 \geq \cdots \geq \sigma_{\min\{m,n\}} = \sigma_{\min}$ called *singular values* of B. If A is not a full rank matrix, then it is often more convenient to use the *reduced singular value decomposition* (RSVD)

$$\mathsf{B} = \widehat{\mathsf{U}}\widehat{\mathsf{S}}\widehat{\mathsf{V}}^T, \tag{2.28}$$

where $\widehat{\mathsf{U}} \in \mathbb{R}^{m\times r}$ and $\widehat{\mathsf{V}} \in \mathbb{R}^{n\times r}$ are matrices with orthonormal columns, $\widehat{\mathsf{S}} \in \mathbb{R}^{r\times r}$ is a nonsingular diagonal matrix with positive diagonal entries $\sigma_1 \geq \cdots \geq \sigma_r = \overline{\sigma}_{\min}$, and $r \leq \min\{m,n\}$ is the rank of B. The matrices $\widehat{\mathsf{U}}$ and $\widehat{\mathsf{V}}$ are formed by the first r columns of U and V. If $\mathbf{x} \in \mathbb{R}^m$, then

$$\mathsf{B}\mathbf{x} = \widehat{\mathsf{U}}\widehat{\mathsf{S}}\widehat{\mathsf{V}}^T\mathbf{x} = (\widehat{\mathsf{U}}\widehat{\mathsf{S}}\widehat{\mathsf{V}}^T)(\widehat{\mathsf{V}}\widehat{\mathsf{S}}\widehat{\mathsf{U}}^T)(\widehat{\mathsf{U}}\widehat{\mathsf{S}}^{-1}\mathsf{V}^T\mathbf{x}) = \mathsf{B}\mathsf{B}^T\mathbf{y},$$

so that

$$\mathrm{Im}\mathsf{B} = \mathrm{Im}\mathsf{B}\mathsf{B}^T. \tag{2.29}$$

If $\mathsf{B} = \widehat{\mathsf{U}}\widehat{\mathsf{S}}\widehat{\mathsf{V}}^T$ is RSVD, then

$$\mathrm{Im}\,\mathsf{B} = \mathrm{Im}\,\widehat{\mathsf{U}}, \quad \mathrm{Ker}\,\mathsf{B} = \left(\mathrm{Im}\,\widehat{\mathsf{V}}\right)^{\perp}.$$

It follows that

$$\mathrm{Im}\,\mathsf{B}^T = (\mathrm{Ker}\,\mathsf{B})^{\perp}. \tag{2.30}$$

The SVD reveals close relations between the properties of a matrix and its singular values. Thus, the rank of $\mathsf{B} \in \mathbb{R}^{m\times n}$ is equal to the number of its nonzero singular values,

$$\|\mathsf{B}\| = \|\mathsf{B}^T\| = \sigma_1, \tag{2.31}$$

and for any vector $\mathbf{x} \in \mathbb{R}^n$

$$\sigma_{\min}\|\mathbf{x}\| \le \|\mathsf{B}\mathbf{x}\| \le \|\mathsf{B}\|\|\mathbf{x}\|. \tag{2.32}$$

Let $\overline{\sigma}_{\min}$ denote the least nonzero singular value of $\mathsf{B} \in \mathbb{R}^{m\times n}$, let $\mathbf{x} \in \mathrm{Im}\mathsf{B}^T$, and consider the RSVD $\mathsf{B} = \widehat{\mathsf{U}}\widehat{\mathsf{S}}\widehat{\mathsf{V}}^T$ with $\widehat{\mathsf{U}} \in \mathbb{R}^{m\times r}$, $\widehat{\mathsf{V}} \in \mathbb{R}^{n\times r}$, and $\widehat{\mathsf{S}} \in \mathbb{R}^{r\times r}$. Then there is $\mathbf{y} \in \mathbb{R}^r$ such that $\mathbf{x} = \widehat{\mathsf{V}}\mathbf{y}$ and

$$\|\mathsf{B}\mathbf{x}\| = \|\widehat{\mathsf{U}}\widehat{\mathsf{S}}\widehat{\mathsf{V}}^T\widehat{\mathsf{V}}\mathbf{y}\| = \|\widehat{\mathsf{U}}\widehat{\mathsf{S}}\mathbf{y}\| = \|\widehat{\mathsf{S}}\mathbf{y}\| \ge \overline{\sigma}_{\min}\|\mathbf{y}\|.$$

Since

$$\|\mathbf{x}\| = \|\widehat{\mathsf{V}}\mathbf{y}\| = \|\mathbf{y}\|,$$

we conclude that

$$\overline{\sigma}_{\min}\|\mathbf{x}\| \le \|\mathsf{B}\mathbf{x}\| \ \text{ for any } \ \mathbf{x} \in \mathrm{Im}\mathsf{B}^T, \tag{2.33}$$

or, equivalently,

$$\overline{\sigma}_{\min}\|\mathbf{x}\| \le \|\mathsf{B}^T\mathbf{x}\| \ \text{ for any } \ \mathbf{x} \in \mathrm{Im}\mathsf{B}. \tag{2.34}$$

The SVD (2.27) can be used to introduce the *Moore–Penrose generalized inverse* of an $m \times n$ matrix B by

$$\mathsf{B}^{\dagger} = \mathsf{V}\mathsf{S}^{\dagger}\mathsf{U}^T,$$

where $\mathsf{S}^{\dagger}$ is the diagonal matrix with the entries $[\mathsf{S}^{\dagger}]_{ii} = 0$ if $\sigma_i = 0$ and $[\mathsf{S}^{\dagger}]_{ii} = \sigma_i^{-1}$ otherwise. It is easy to check that

$$\mathsf{B}\mathsf{B}^{\dagger}\mathsf{B} = \mathsf{U}\mathsf{S}\mathsf{V}^T\mathsf{V}\mathsf{S}^{\dagger}\mathsf{U}^T\mathsf{U}\mathsf{S}\mathsf{V}^T = \mathsf{U}\mathsf{S}\mathsf{V}^T = \mathsf{B}, \tag{2.35}$$

so that the Moore–Penrose generalized inverse is a generalized inverse. If B is a full row rank matrix, then it may be checked directly that

$$\mathsf{B}^{\dagger} = \mathsf{B}^T(\mathsf{B}\mathsf{B}^T)^{-1}.$$

If B is a singular matrix and $\mathbf{c} \in \mathrm{Im}\mathsf{B}$, then $\mathbf{x}_{\mathrm{LS}} = \mathsf{B}^{\dagger}\mathbf{c}$ is a solution of the system of linear equations $\mathsf{B}\mathbf{x} = \mathbf{c}$, i.e.,

$$\mathsf{B}\mathbf{x}_{\mathrm{LS}} = \mathbf{c}.$$

Notice that $\mathbf{x}_{\mathrm{LS}} \in \mathrm{Im}\mathsf{B}^T$, so that if $\overline{\mathbf{x}}$ is any other solution, then $\overline{\mathbf{x}} = \mathbf{x}_{\mathrm{LS}} + \mathbf{d}$, where $\mathbf{d} \in \mathrm{Ker}\mathsf{B}$, $\mathbf{x}_{\mathrm{LS}}^T\mathbf{d} = 0$, and

$$\|\mathbf{x}_{\mathrm{LS}}\|^2 \leq \|\overline{\mathbf{x}}_{\mathrm{LS}}\|^2 + \|\mathbf{d}\|^2 = \|\overline{\mathbf{x}}\|^2. \tag{2.36}$$

The vector $\mathbf{x}_{\mathrm{LS}}$ is called the *least square solution* of $\mathsf{B}\mathbf{x} = \mathbf{c}$.

Obviously

$$\|\mathsf{B}^{\dagger}\| = \overline{\sigma}_{\min}^{-1}, \tag{2.37}$$

where $\overline{\sigma}_{\min}$ denotes the least nonzero singular value of B, so that

$$\|\mathbf{x}_{\mathrm{LS}}\| = \|\mathsf{B}^{\dagger}\mathbf{c}\| \leq \overline{\sigma}_{\min}^{-1}\|\mathbf{c}\|. \tag{2.38}$$

It can be verified directly that

$$\left(\mathsf{B}^{\dagger}\right)^T = \left(\mathsf{B}^T\right)^{\dagger}.$$

2.9 Graphs, Walks, and Adjacency Matrices

We shall need some simple results of graph theory and linear algebra. Let us recall that the *vertices* V_i and the *edges* of the *graph of the mesh of the triangulation* $\mathscr{T} = \{\tau_i\}$ of a polyhedral domain Ω are the nodes of the mesh and their adjacent couples V_iV_j, respectively. Recall that the edges V_i and V_j are *adjacent* if there is an element $\tau_k \in \mathscr{T}$ such that V_iV_j is the edge of τ_k. The graph is fully described by the *adjacency matrix* D with the nonzero entries d_{ij} equal to one if the nodes V_i and V_j are adjacent. Since the graph of the mesh is not oriented and does not contain loops, the adjacency matrix is symmetric and its diagonal entries d_{ii} are equal to zero. Let us also recall that the *walk of length k* in the mesh of $\mathscr{T}$ is a sequence of the distinct nodes $V_{i_1}, \ldots, V_{i_k}$ such that the edges $V_{i_j}V_{i_{j+1}}$, $j = 1, 2, \ldots, k-1$ belong to the graph of the mesh. Thus

$$d_{i_j i_{j+1}} = 1, \quad j = 1, \ldots, k-1.$$

The walk $(V_{i_1}, \ldots, V_{i_k})$ starts at V_{i_1} and ends at V_{i_k}. Moreover, we call a walk between nodes V_i and V_k an (i, k)-*walk*. We use the following well-known observation [5].

Lemma 2.3 *Let* D *be the adjacency matrix of the mesh of* $\mathscr{T}$ *and* $\mathsf{B} = \mathsf{D}^k$. *Then each entry* b_{ij} *of* B *gives the number of distinct* (i, j)*-walks of length* k.

Proof To see why our lemma holds, we use induction on k. For $k = 1$ the claim follows immediately from the definition of D. Suppose that for some $k \geq 1$ the entry b_{ij} in $\mathsf{B} = \mathsf{D}^k$ gives the number of distinct (i, j)-walks of length k. For convenience we denote $\mathsf{C} = \mathsf{D}^{k+1}$, so that $\mathsf{C} = \mathsf{B}\mathsf{D}$. The entries of C are given by

$$c_{ij} = \sum_{\ell=1}^{n} b_{i\ell} d_{\ell j},$$

where the number $b_{i\ell}$ gives the number of distinct (i, ℓ)-walks of length k and $d_{\ell j} = 0$ or $d_{\ell j} = 1$. If a particular edge $V_\ell V_j$ is not in the mesh (graph) then $d_{\ell j} = 0$ and $b_{i\ell} d_{\ell j} = 0$. Thus, there is no (i, j)-walk of length $k + 1$ with the last-but-one node ℓ. On the other hand, if $d_{\ell, j} = 1$ then we can prolong each (i, ℓ)-walk to an (i, j)-walk. Thus c_{ij} gives the number of all distinct (i, j)-walks of length $k + 1$. □

The following corollary follows easily from Lemma 2.3.

Corollary 2.1 *Let* D *denote the adjacency matrix of a given mesh and* $\mathbf{e} = [e_i]$, $e_i = 1$, $i = 1, 2, \ldots, n$. *Then the number* $w(i, k)$ *of distinct walks of length* k *starting at node* i *is given by*

$$w(i, k) = [\mathsf{D}^k \mathbf{e}]_i.$$

If the mesh is approximately regular, we expect that more walks of length k originate from the nodes that are near a center of the mesh than from the vertices that are far from it. It simply follows that the node with the index i which satisfies $w(i, k) \geq w(j, k)$, $j = 1, 2, \ldots, n$, for sufficiently large k is in a sense near to the center of the mesh.

References

1. Demmel, J.W.: Applied Numerical Linear Algebra. SIAM, Philadelphia (1997)
2. Laub, A.J.: Matrix Analysis for Scientists and Engineers. SIAM, Philadelphia (2005)
3. Golub, G.H., Van Loan, C.F.: Matrix Computations, 2nd edn. Johns Hopkins University Press, Baltimore (1989)
4. Dostál, Z., Kozubek, T., Markopoulos, A., Menšík, M.: Cholesky factorization of a positive semidefinite matrix with known kernel. Appl. Math. Comput. **217**, 6067–6077 (2011)
5. Diestel, R.: Graph Theory. Springer, Heidelberg (2005)

Chapter 3
Optimization

In this chapter, we briefly review the results concerning the minimization of quadratic functions to the extent which is sufficient for understanding the algorithms described in Part II. The results are presented with specialized arguments, typically algebraic, that exploit the specific structure of these problems. Systematic exposition of optimization theory in the framework of nonlinear optimization can be found in the books by Bertsekas [1], Nocedal and Wright [2], Conn, Gould, and Toint [3], Bazaraa, Sherali, and Shetty [4], or Griva, Nash, and Sofer [5].

3.1 Optimization Problems and Solutions

Optimization problems considered in this book are described by a *cost (objective, target) function* f defined on a subset $\mathcal{D} \subseteq \mathbb{R}^n$ and by a *constraint set* $\Omega \subseteq \mathcal{D}$. The elements of Ω are called *feasible vectors*. Important ingredients of scalable algorithms for the frictionless contact problems are efficient algorithms for the solution of *quadratic programming (QP) problems* with a quadratic cost function f and a constraint set $\Omega \subseteq \mathbb{R}^n$ described by linear equalities and inequalities. The solution of problems with friction requires effective algorithms for special *quadratically constrained quadratic programmes (QCQP)* with the constraints described by linear equalities and separable quadratic inequalities.

We look either for a solution $\overline{\mathbf{x}} \in \mathbb{R}^n$ of the *unconstrained minimization problem* which satisfies

$$f(\overline{\mathbf{x}}) \leq f(\mathbf{x}), \ \mathbf{x} \in \mathbb{R}^n, \tag{3.1}$$

or for a solution $\overline{\mathbf{x}} \in \Omega$ of the *constrained minimization problem*

$$f(\overline{\mathbf{x}}) \leq f(\mathbf{x}), \quad \mathbf{x} \in \Omega, \quad \Omega \subset \mathbb{R}^n. \tag{3.2}$$

A solution of the minimization problem is called its *minimizer* or *global minimizer*.

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3_3

A nonzero vector $\mathbf{d} \in \mathbb{R}^n$ is a *feasible direction* of Ω at a feasible point $\mathbf{x}$ if $\mathbf{x} + \varepsilon\mathbf{d} \in \Omega$ for all sufficiently small $\varepsilon > 0$. A nonzero vector $\mathbf{d} \in \mathbb{R}^n$ is a *recession direction*, or simply a *direction*, of Ω if for each $\mathbf{x} \in \Omega$, $\mathbf{x} + \alpha\mathbf{d} \in \Omega$ for all $\alpha > 0$.

3.2 Unconstrained Quadratic Programming

Let us first recall some simple results which concern unconstrained quadratic programming.

3.2.1 Quadratic Cost Functions

We consider the cost functions in the form

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathsf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x}, \tag{3.3}$$

where $\mathsf{A} \in \mathbb{R}^{n\times n}$ denotes a given SPS or SPD matrix of order n and $\mathbf{b} \in \mathbb{R}^n$.

If $\mathbf{x}, \mathbf{d} \in \mathbb{R}^n$, then using elementary computations and $\mathsf{A} = \mathsf{A}^T$, we get

$$f(\mathbf{x} + \mathbf{d}) = f(\mathbf{x}) + (\mathsf{A}\mathbf{x} - \mathbf{b})^T\mathbf{d} + \frac{1}{2}\mathbf{d}^T\mathsf{A}\mathbf{d}. \tag{3.4}$$

The formula (3.4) is *Taylor's expansion* of f at $\mathbf{x}$, so that the *gradient* of f at $\mathbf{x}$ is given by

$$\nabla f(\mathbf{x}) = \mathsf{A}\mathbf{x} - \mathbf{b}, \tag{3.5}$$

and the *Hessian* of f at $\mathbf{x}$ is given by

$$\nabla^2 f(\mathbf{x}) = \mathsf{A}.$$

Taylor's expansion will be our simple but powerful tool in what follows.

A vector $\mathbf{d}$ is a *decrease direction* of f at $\mathbf{x}$ if

$$f(\mathbf{x} + \varepsilon\mathbf{d}) < f(\mathbf{x})$$

for all sufficiently small values of $\varepsilon > 0$. Using Taylor's expansion (3.4) in the form

$$f(\mathbf{x} + \varepsilon\mathbf{d}) = f(\mathbf{x}) + \varepsilon(\mathsf{A}\mathbf{x} - \mathbf{b})^T\mathbf{d} + \frac{\varepsilon^2}{2}\mathbf{d}^T\mathsf{A}\mathbf{d},$$

we get that $\mathbf{d}$ is a decrease direction if and only if

$$(\mathsf{A}\mathbf{x} - \mathbf{b})^T\mathbf{d} < 0.$$

3.2.2 Unconstrained Minimization of Quadratic Functions

The following proposition gives algebraic conditions that are satisfied by the solutions of the unconstrained QP problem to find

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}), \tag{3.6}$$

where f is a quadratic function defined by (3.3).

Proposition 3.1 *Let the quadratic function f be defined by an SPS or SPD matrix $\mathsf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{b} \in \mathbb{R}^n$. Then the following statements hold:*
(i) A vector $\overline{\mathbf{x}}$ is a solution of the unconstrained minimization problem (3.6) if and only if

$$\nabla f(\overline{\mathbf{x}}) = \mathsf{A}\overline{\mathbf{x}} - \mathbf{b} = \mathbf{o}. \tag{3.7}$$

(ii) The minimization problem (3.6) has a unique solution if and only if A is SPD.

Proof The proof is a simple corollary of Taylor's expansion formula (3.4). □

Remark 3.1 Condition (3.7) can be written as a variational equality

$$(\mathsf{A}\overline{\mathbf{x}})^T(\mathbf{x} - \overline{\mathbf{x}}) = \mathbf{b}^T(\mathbf{x} - \overline{\mathbf{x}}), \quad \mathbf{x} \in \mathbb{R}^n.$$

Examining the gradient condition (3.7), we get that problem (3.6) has a solution if and only if A is SPS and

$$\mathbf{b} \in \mathrm{Im}\mathsf{A}. \tag{3.8}$$

Denoting by R a matrix the columns of which span $\mathrm{Ker}\mathsf{A}$, we can rewrite (3.8) as $\mathsf{R}^T\mathbf{b} = \mathbf{o}$. This condition has a simple mechanical interpretation: if a mechanical system is in equilibrium, the external forces must be orthogonal to the rigid body motions.

If $\mathbf{b} \in \mathrm{Im}\mathsf{A}$, a solution of (3.6) is given by

$$\overline{\mathbf{x}} = \mathsf{A}^+\mathbf{b},$$

where A^+ is a left generalized inverse introduced in Sect. 2.3. After substituting into f and simple manipulations, we get

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = -\frac{1}{2}\mathbf{b}^T\mathsf{A}^+\mathbf{b}. \tag{3.9}$$

In particular, if A is positive definite, then

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = -\frac{1}{2}\mathbf{b}^T\mathsf{A}^{-1}\mathbf{b}. \tag{3.10}$$

The above formulae can be used to develop useful estimates. Indeed, if (3.8) holds and $\mathbf{x} \in \mathbb{R}^n$, we get

$$f(\mathbf{x}) \geq -\frac{1}{2}\mathbf{b}^T \mathsf{A}^+ \mathbf{b} = -\frac{1}{2}\mathbf{b}^T \mathsf{A}^\dagger \mathbf{b} \geq -\frac{1}{2}\|\mathsf{A}^\dagger\|\|\mathbf{b}\|^2 = -\frac{\|\mathbf{b}\|^2}{2\bar{\lambda}_{\min}},$$

where $\mathsf{A}^\dagger$ denotes the Moore–Penrose generalized inverse and $\bar{\lambda}_{\min}$ denotes the least nonzero eigenvalue of A. In particular, it follows that if A is positive definite and $\lambda_{\min}$ denotes the least eigenvalue of A, then for any $\mathbf{x} \in \mathbb{R}^n$

$$f(\mathbf{x}) \geq -\frac{1}{2}\mathbf{b}^T \mathsf{A}^{-1} \mathbf{b} \geq -\frac{1}{2}\|\mathsf{A}^{-1}\|\|\mathbf{b}\|^2 = -\frac{\|\mathbf{b}\|^2}{2\lambda_{\min}}. \tag{3.11}$$

If the dimension n of the unconstrained minimization problem (3.6) is large, then it can be too ambitious to look for a solution which satisfies the gradient condition (3.7) exactly. A natural idea is to consider the weaker condition

$$\|\nabla f(\mathbf{x})\| \leq \varepsilon \tag{3.12}$$

with a small epsilon. If $\mathbf{x}$ satisfies the latter condition with ε sufficiently small and A nonsingular, then $\mathbf{x}$ is near the unique solution $\widehat{\mathbf{x}}$ as

$$\|\mathbf{x} - \widehat{\mathbf{x}}\| = \|\mathsf{A}^{-1}\mathsf{A}\left(\mathbf{x} - \widehat{\mathbf{x}}\right)\| = \|\mathsf{A}^{-1}(\mathsf{A}\mathbf{x} - \mathbf{b})\| \leq \|\mathsf{A}^{-1}\|\|\nabla f(\mathbf{x})\|. \tag{3.13}$$

The typical "solution" returned by an iterative solver is just $\mathbf{x}$ that satisfies (3.12).

3.3 Convexity

Intuitively, convexity is a property of the sets that contain the joining segment with any two points. More formally, a subset Ω of $\mathbb{R}^n$ is *convex* if for any $\mathbf{x}$ and $\mathbf{y}$ in Ω and $\alpha \in (0, 1)$, the vector $\mathbf{s} = \alpha\mathbf{x} + (1 - \alpha)\mathbf{y}$ is also in Ω.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_k$ be vectors of $\mathbb{R}^n$. If $\alpha_1, \ldots, \alpha_k$ are scalars such that

$$\alpha_i \geq 0, \quad i = 1, \ldots, k, \qquad \sum_{i=1}^{k} \alpha_i = 1,$$

then the vector $\mathbf{v} = \sum_{i=1}^{k} \alpha_i \mathbf{x}_i$ is said to be a *convex combination* of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$. The *convex hull* of $\mathbf{x}_1, \ldots, \mathbf{x}_k$, denoted $\mathrm{Conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$, is the set of all convex combinations of $\mathbf{x}_1, \ldots, \mathbf{x}_k$.

3.3.1 Convex Quadratic Functions

Given a convex set $\Omega \in \mathbb{R}^n$, a mapping $h : \Omega \to \mathbb{R}$ is said to be a *convex function* if its epigraph is convex, that is, if

$$h\left(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}\right) \leq \alpha h(\mathbf{x}) + (1-\alpha)h(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in \Omega$ and $\alpha \in (0, 1)$, and it is *strictly convex* if

$$h\big(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}\big) < \alpha h(\mathbf{x}) + (1-\alpha)h(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in \Omega,\ \mathbf{x} \neq \mathbf{y}$, and $\alpha \in (0, 1)$.

The following proposition gives a characterization of convex functions.

Proposition 3.2 *Let V be a subspace of $\mathbb{R}^n$. The restriction $f|V$ of a quadratic function f with the Hessian matrix A to V is convex if and only if $\mathsf{A}|V$ is positive semidefinite, and $f|V$ is strictly convex if and only if $\mathsf{A}|V$ is positive definite.*

Proof Let V be a subspace, let $\mathbf{x}, \mathbf{y} \in V, \alpha \in (0, 1)$, and $\mathbf{s} = \alpha\mathbf{x} + (1-\alpha)\mathbf{y}$. Then by Taylor's expansion (3.4) of f at $\mathbf{s}$

$$\begin{aligned} f(\mathbf{s}) + \nabla f(\mathbf{s})^T(\mathbf{x}-\mathbf{s}) + \frac{1}{2}(\mathbf{x}-\mathbf{s})^T\mathsf{A}(\mathbf{x}-\mathbf{s}) &= f(\mathbf{x}), \\ f(\mathbf{s}) + \nabla f(\mathbf{s})^T(\mathbf{y}-\mathbf{s}) + \frac{1}{2}(\mathbf{y}-\mathbf{s})^T\mathsf{A}(\mathbf{y}-\mathbf{s}) &= f(\mathbf{y}). \end{aligned}$$

Multiplying the first equation by α, the second equation by $1-\alpha$, and summing up, we get

$$\begin{aligned} f(\mathbf{s}) + \frac{\alpha}{2}(\mathbf{x}-\mathbf{s})^T\mathsf{A}(\mathbf{x}-\mathbf{s}) + \frac{1-\alpha}{2}(\mathbf{y}-\mathbf{s})^T\mathsf{A}(\mathbf{y}-\mathbf{s}) \\ = \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}). \end{aligned} \tag{3.14}$$

It follows that if $\mathsf{A}|V$ is positive semidefinite, then $f|V$ is convex. Moreover, since $\mathbf{x} = \mathbf{y}$ is equivalent to $\mathbf{x} = \mathbf{s}$ and $\mathbf{y} = \mathbf{s}$, it follows that if $\mathsf{A}|V$ is positive definite, then $f|V$ is strictly convex.

Let us now assume that $f|V$ is convex, let $\mathbf{z} \in V$, $\alpha = \frac{1}{2}$, and denote $\mathbf{x} = 2\mathbf{z}$, $\mathbf{y} = \mathbf{o}$. Then $\mathbf{s} = \mathbf{z},\ \mathbf{x} - \mathbf{s} = \mathbf{z},\ \mathbf{y} - \mathbf{s} = -\mathbf{z}$, and substituting into (3.14) results in

$$f(\mathbf{s}) + \frac{1}{2}\mathbf{z}^T\mathsf{A}\mathbf{z} = \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}).$$

Since $\mathbf{z} \in V$ is arbitrary and $f|V$ is assumed to be convex, it follows that

$$\frac{1}{2}\mathbf{z}^T\mathsf{A}\mathbf{z} = \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}) - f\left(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}\right) \geq 0.$$

Thus $\mathsf{A}|V$ is positive semidefinite. Moreover, if $f|V$ is strictly convex, then $\mathsf{A}|V$ is positive definite. □

The strictly convex quadratic functions have a nice property that $f(\mathbf{x}) \to \infty$ when $\|\mathbf{x}\| \to \infty$. The functions with this property are called *coercive functions*. More generally, a function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *coercive on* $\Omega \subseteq \mathbb{R}^n$ if

$$f(\mathbf{x}) \to \infty \quad \text{for} \quad \|\mathbf{x}\| \to \infty, \ \mathbf{x} \in \Omega.$$

3.3.2 Minimizers of Convex Function

Under the convexity assumptions, each local minimizer is a global minimizer. We shall formulate this result together with some observations concerning the set of solutions.

Proposition 3.3 *Let f and $\Omega \subseteq \mathbb{R}^n$ be a convex quadratic function defined by (3.3) and a closed convex set, respectively. Then the following statements hold:*
(i) Each local minimizer of f subject to $\mathbf{x} \in \Omega$ is a global minimizer of f subject to $\mathbf{x} \in \Omega$.
(ii) If $\overline{\mathbf{x}}$, $\overline{\mathbf{y}}$ are two minimizers of f subject to $\mathbf{x} \in \Omega$, then

$$\overline{\mathbf{x}} - \overline{\mathbf{y}} \in \mathrm{Ker}\mathsf{A} \cap \mathrm{Span}\{\mathbf{b}\}^{\perp}.$$

(iii) If f is strictly convex on Ω and $\overline{\mathbf{x}}$, $\overline{\mathbf{y}}$ are two minimizers of f subject to $\mathbf{x} \in \Omega$, then $\overline{\mathbf{x}} = \overline{\mathbf{y}}$.

Proof (i) Let $\overline{\mathbf{x}} \in \Omega$ and $\overline{\mathbf{y}} \in \Omega$ be local minimizers of f subject to $\mathbf{x} \in \Omega$, $f(\overline{\mathbf{x}}) < f(\overline{\mathbf{y}})$. Denoting $\mathbf{y}_\alpha = \alpha\overline{\mathbf{x}} + (1-\alpha)\overline{\mathbf{y}}$ and using that f is convex, we get

$$f(\mathbf{y}_\alpha) = f(\alpha\overline{\mathbf{x}} + (1-\alpha)\overline{\mathbf{y}}) \le \alpha f(\overline{\mathbf{x}}) + (1-\alpha) f(\overline{\mathbf{y}}) < f(\overline{\mathbf{y}})$$

for every $\alpha \in (0, 1)$. Since

$$\|\overline{\mathbf{y}} - \mathbf{y}_\alpha\| = \alpha\|\overline{\mathbf{y}} - \overline{\mathbf{x}}\|,$$

the inequality contradicts the assumption that $\overline{\mathbf{y}}$ is a local minimizer.
(ii) Let $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ be global minimizers of f on Ω. Then for any $\alpha \in [0, 1]$

$$\overline{\mathbf{x}} + \alpha(\overline{\mathbf{y}} - \overline{\mathbf{x}}) = (1-\alpha)\overline{\mathbf{x}} + \alpha\overline{\mathbf{y}} \in \Omega, \quad \overline{\mathbf{y}} + \alpha(\overline{\mathbf{x}} - \overline{\mathbf{y}}) = (1-\alpha)\overline{\mathbf{y}} + \alpha\overline{\mathbf{x}} \in \Omega.$$

Moreover, using Taylor's formula, we get

$$0 \le f\big(\overline{\mathbf{x}} + \alpha(\overline{\mathbf{y}} - \overline{\mathbf{x}})\big) - f(\overline{\mathbf{x}}) = \alpha(\mathsf{A}\overline{\mathbf{x}} - \mathbf{b})^T(\overline{\mathbf{y}} - \overline{\mathbf{x}}) + \frac{\alpha^2}{2}(\overline{\mathbf{y}} - \overline{\mathbf{x}})^T\mathsf{A}(\overline{\mathbf{y}} - \overline{\mathbf{x}}),$$
$$0 \le f\big(\overline{\mathbf{y}} + \alpha(\overline{\mathbf{x}} - \overline{\mathbf{y}})\big) - f(\overline{\mathbf{y}}) = \alpha(\mathsf{A}\overline{\mathbf{y}} - \mathbf{b})^T(\overline{\mathbf{x}} - \overline{\mathbf{y}}) + \frac{\alpha^2}{2}(\overline{\mathbf{x}} - \overline{\mathbf{y}})^T\mathsf{A}(\overline{\mathbf{x}} - \overline{\mathbf{y}}).$$

Since the latter inequalities hold for arbitrarily small α, it follows that

$$(\mathsf{A}\overline{\mathbf{x}} - \mathbf{b})^T(\overline{\mathbf{y}} - \overline{\mathbf{x}}) \ge 0 \quad \text{and} \quad (\mathsf{A}\overline{\mathbf{y}} - \mathbf{b})^T(\overline{\mathbf{x}} - \overline{\mathbf{y}}) \ge 0.$$

After summing up the latter inequalities and simple manipulations, we have

$$-(\overline{\mathbf{x}} - \overline{\mathbf{y}})^T\mathsf{A}(\overline{\mathbf{x}} - \overline{\mathbf{y}}) \ge 0.$$

Since the convexity of f implies by Proposition 3.2 that A is positive semidefinite, it follows that $\overline{\mathbf{x}} - \overline{\mathbf{y}} \in \mathrm{Ker}\mathsf{A}$.

If $f(\overline{\mathbf{x}}) = f(\overline{\mathbf{y}})$ and $\overline{\mathbf{x}}, \overline{\mathbf{y}} \in \mathrm{Ker}\mathsf{A}$, then

$$\mathbf{b}^T(\overline{\mathbf{x}} - \overline{\mathbf{y}}) = f(\overline{\mathbf{x}}) - f(\overline{\mathbf{y}}) = 0,$$

i.e., $\overline{\mathbf{x}} - \overline{\mathbf{y}} \in \mathrm{Span}\{\mathbf{b}\}^{\perp}$.

(iii) Let f be strictly convex and let $\overline{\mathbf{x}}, \overline{\mathbf{y}} \in \Omega$ be different global minimizers of f on Ω, so that $f(\overline{\mathbf{x}}) = f(\overline{\mathbf{y}})$. Then $\mathrm{Ker}\mathsf{A} = \{\mathbf{o}\}$ and by (ii) $\overline{\mathbf{x}} - \overline{\mathbf{y}} = \mathbf{o}$. □

3.3.3 Existence of Minimizers

Since quadratic functions are continuous, existence of at least one minimizer is guaranteed by the Weierstrass theorem provided Ω is compact, that is, closed and bounded. The following standard results do not assume that Ω is bounded.

Proposition 3.4 *Let f be a convex quadratic function and let Ω denote a closed convex set. Then the following statements hold:*
(i) If f is strictly convex, then there is a unique minimizer of f subject to $\mathbf{x} \in \Omega$.
(ii) If f is coercive on Ω, then a global minimizer of f subject to $\mathbf{x} \in \Omega$ exists.
(iii) If f is bounded from below on Ω, then there is a global minimizer of f subject to $\mathbf{x} \in \Omega$.

Proof (i) If f is strictly convex, it follows by Proposition 3.2 that A is SPD and $\mathbf{z} = \mathsf{A}^{-1}\mathbf{b}$ is by Proposition 3.1 the unique minimizer of f on $\mathbb{R}^n$. Thus for any $\mathbf{x} \in \mathbb{R}^n$

$$f(\mathbf{x}) \ge f(\mathbf{z}).$$

It follows that the infimum of $f(\mathbf{x})$ subject to $\mathbf{x} \in \Omega$ exists, and there is a sequence of vectors $\mathbf{x}^k \in \Omega$ such that

$$\lim_{k\to\infty} f(\mathbf{x}^k) = \inf_{\mathbf{x}\in\Omega} f(\mathbf{x}).$$

The sequence $\{\mathbf{x}^k\}$ is bounded as

$$f(\mathbf{x}^k) - f(\mathbf{z}) = \frac{1}{2}(\mathbf{x}^k - \mathbf{z})^T \mathsf{A}(\mathbf{x}^k - \mathbf{z}) \geq \frac{\lambda_{\min}}{2}\|\mathbf{x}^k - \mathbf{z}\|^2,$$

where $\lambda_{\min}$ denotes the least eigenvalue of A. It follows that $\{\mathbf{x}^k\}$ has at least one cluster point $\overline{\mathbf{x}} \in \Omega$. Since f is continuous, we get

$$f(\overline{\mathbf{x}}) = \inf_{\mathbf{x}\in\Omega} f(\mathbf{x}).$$

The uniqueness follows by Proposition 3.3.
(ii) The proof is similar to that of (i). See, e.g., Bertsekas [1, Proposition A.8].
(iii) The statement is the well-known Frank–Wolfe theorem [6]. □

3.3.4 *Projections to Convex Sets*

Let us define the *projection* P_Ω *to the (closed) convex set* $\Omega \subset \mathbb{R}^n$ as a mapping which assigns to each $\mathbf{x} \in \mathbb{R}^n$ its nearest vector $\widehat{\mathbf{x}} \in \Omega$ as in Fig. 3.1. The following proposition concerns the projection induced by the Euclidean scalar product.

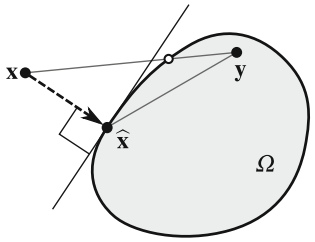

Fig. 3.1 Projection to the convex set

Proposition 3.5 *Let* $\Omega \subseteq \mathbb{R}^n$ *be a nonempty closed convex set and* $\mathbf{x} \in \mathbb{R}^n$. *Then there is a unique point* $\widehat{\mathbf{x}} \in \Omega$ *with the minimum Euclidean distance from* $\mathbf{x}$, *and for any* $\mathbf{y} \in \Omega$

$$(\mathbf{x} - \widehat{\mathbf{x}})^T(\mathbf{y} - \widehat{\mathbf{x}}) \leq 0. \tag{3.15}$$

Proof Since the proof is trivial for $\mathbf{x} \in \Omega$, let us assume that $\mathbf{x} \notin \Omega$ is arbitrary but fixed and observe that the function f defined on $\mathbb{R}^n$ by

$$f(\mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2 = \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{x} + \|\mathbf{x}\|^2$$

has the Hessian

$$\nabla^2 f(\mathbf{y}) = 2\mathrm{I}.$$

The identity matrix being positive definite, it follows by Proposition 3.2 that f is strictly convex, so that the unique minimizer $\widehat{\mathbf{x}} \in \Omega$ of $f(\mathbf{y})$ subject to $\mathbf{y} \in \Omega$ exists by Proposition 3.4(i).

If $\mathbf{y} \in \Omega$ and $\alpha \in (0, 1)$, then by convexity of Ω

$$(1 - \alpha)\widehat{\mathbf{x}} + \alpha\mathbf{y} = \widehat{\mathbf{x}} + \alpha(\mathbf{y} - \widehat{\mathbf{x}}) \in \Omega,$$

so that for any $\mathbf{x} \in \mathbb{R}^n$

$$\|\mathbf{x} - \widehat{\mathbf{x}}\|^2 \leq \|\mathbf{x} - \widehat{\mathbf{x}} - \alpha(\mathbf{y} - \widehat{\mathbf{x}})\|^2.$$

Using simple manipulations and the latter inequality, we get

$$\begin{aligned}\|\mathbf{x} - \widehat{\mathbf{x}} - \alpha(\mathbf{y} - \widehat{\mathbf{x}})\|^2 &= \|\widehat{\mathbf{x}} - \mathbf{x}\|^2 + \alpha^2\|\mathbf{y} - \widehat{\mathbf{x}}\|^2 - 2\alpha(\mathbf{x} - \widehat{\mathbf{x}})^T(\mathbf{y} - \widehat{\mathbf{x}}) \\ &\leq \|\mathbf{x} - \widehat{\mathbf{x}} - \alpha(\mathbf{y} - \widehat{\mathbf{x}})\|^2 \\ &\quad + \alpha^2\|\mathbf{y} - \widehat{\mathbf{x}}\|^2 - 2\alpha(\mathbf{x} - \widehat{\mathbf{x}})^T(\mathbf{y} - \widehat{\mathbf{x}}).\end{aligned}$$

Thus

$$2\alpha(\mathbf{x} - \widehat{\mathbf{x}})^T(\mathbf{y} - \widehat{\mathbf{x}}) \leq \alpha^2\|\mathbf{y} - \widehat{\mathbf{x}}\|^2$$

for any $\alpha \in (0, 1)$. To obtain (3.15), just divide the last inequality by α and observe that α may be arbitrarily small. □

Using Proposition 3.5, it is not difficult to show that the mapping P_Ω which assigns to each $\mathbf{x} \in \mathbb{R}^n$ its projection to Ω is *nonexpansive* as in Fig. 3.2.

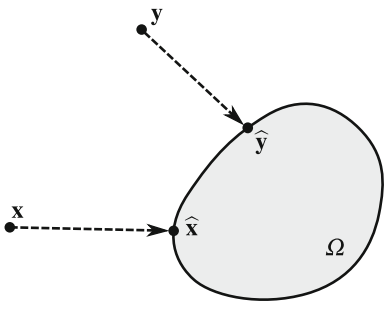

Fig. 3.2 Projection P_Ω is nonexpansive

Corollary 3.1 *Let $\Omega \subseteq \mathbb{R}^n$ be a nonempty closed convex set, and for any $\mathbf{x} \in \mathbb{R}^n$, let $\widehat{\mathbf{x}} \in \Omega$ denote the projection of $\mathbf{x}$ to Ω. Then for any $\mathbf{x}, \mathbf{y} \in \Omega$*

$$\|\widehat{\mathbf{x}} - \widehat{\mathbf{y}}\| \leq \|\mathbf{x} - \mathbf{y}\|. \tag{3.16}$$

Proof If $\mathbf{x}, \mathbf{y} \in \mathbb{R}$, then by Proposition 3.5 their projections $\widehat{\mathbf{x}}, \widehat{\mathbf{y}}$ to Ω satisfy

$$(\mathbf{x} - \widehat{\mathbf{x}})^T (\mathbf{z} - \widehat{\mathbf{x}}) \leq 0 \quad \text{and} \quad (\mathbf{y} - \widehat{\mathbf{y}})^T (\mathbf{z} - \widehat{\mathbf{y}}) \leq 0$$

for any $\mathbf{z} \in \Omega$. Substituting $\mathbf{z} = \widehat{\mathbf{y}}$ into the first inequality, $\mathbf{z} = \widehat{\mathbf{x}}$ into the second inequality, and summing up, we get

$$(\mathbf{x} - \widehat{\mathbf{x}} - \mathbf{y} + \widehat{\mathbf{y}})^T (\widehat{\mathbf{y}} - \widehat{\mathbf{x}}) \leq 0.$$

After rearranging the entries and using the Schwarz inequality, we get

$$\|\widehat{\mathbf{x}} - \widehat{\mathbf{y}}\|^2 \leq (\mathbf{x} - \mathbf{y})^T (\widehat{\mathbf{x}} - \widehat{\mathbf{y}}) \leq \|\mathbf{x} - \mathbf{y}\| \|\widehat{\mathbf{x}} - \widehat{\mathbf{y}}\|,$$

which proves (3.16). □

3.4 Equality Constrained Problems

We shall now consider the problems with the constraint set described by a set of linear equations. More formally, we shall look for

$$\min_{\mathbf{x} \in \Omega_E} f(\mathbf{x}), \tag{3.17}$$

where f is a convex quadratic function defined by (3.3), $\Omega_E = \{\mathbf{x} \in \mathbb{R}^n : \mathsf{B}\mathbf{x} = \mathbf{c}\}$, $\mathsf{B} \in \mathbb{R}^{m \times n}$, and $\mathbf{c} \in \mathrm{Im}\mathsf{B}$. We assume that $\mathsf{B} \neq \mathsf{O}$ is not a full column rank matrix, so that $\mathrm{Ker}\mathsf{B} \neq \{\mathbf{o}\}$, but we admit dependent rows of B. It is easy to check that Ω_E is a nonempty closed convex set.

A feasible set Ω_E is a *linear manifold* of the form

$$\Omega_E = \overline{\mathbf{x}} + \mathrm{Ker}\mathsf{B},$$

where $\overline{\mathbf{x}}$ is any vector which satisfies

$$\mathsf{B}\overline{\mathbf{x}} = \mathbf{c}.$$

Thus, a nonzero vector $\mathbf{d} \in \mathbb{R}^n$ is a feasible direction of Ω_E at any $\mathbf{x} \in \Omega_E$ if and only if $\mathbf{d} \in \mathrm{Ker}\mathsf{B}$, and $\mathbf{d}$ is a recession direction of Ω_E if and only if $\mathbf{d} \in \mathrm{Ker}\mathsf{B}$.

Substituting $\mathbf{x} = \overline{\mathbf{x}} + \mathbf{z},\ \mathbf{z} \in \mathrm{Ker}\mathsf{B}$, we can reduce (3.17) to the minimization of

$$f_{\overline{x}}(\mathbf{z}) = \frac{1}{2}\mathbf{z}^T \mathsf{A}\mathbf{z} - (\mathbf{b} - \mathsf{A}\overline{\mathbf{x}})^T \mathbf{z} \tag{3.18}$$

over the subspace $\mathrm{Ker}\mathsf{B}$. Thus we can assume, without loss of generality, that $\mathbf{c} = \mathbf{o}$ in the definition of Ω_E. We shall occasionally use this assumption to simplify our exposition.

A useful tool for the analysis of equality constrained problems is the *Lagrangian function* $L_0 : \mathbb{R}^{n+m} \to \mathbb{R}$ defined by

$$L_0(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T(\mathsf{B}\mathbf{x} - \mathbf{c}) = \frac{1}{2}\mathbf{x}^T \mathsf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x} + (\mathsf{B}\mathbf{x} - \mathbf{c})^T\boldsymbol{\lambda}. \tag{3.19}$$

Obviously

$$\nabla^2_{\mathbf{xx}} L_0(\mathbf{x}, \boldsymbol{\lambda}) = \nabla^2 f(\mathbf{x}) = \mathsf{A}, \tag{3.20}$$

$$\nabla_{\mathbf{x}} L_0(\mathbf{x}, \boldsymbol{\lambda}) = \nabla f(\mathbf{x}) + \mathsf{B}^T\boldsymbol{\lambda} = \mathsf{A}\mathbf{x} - \mathbf{b} + \mathsf{B}^T\boldsymbol{\lambda}, \tag{3.21}$$

$$L_0(\mathbf{x} + \mathbf{d}, \boldsymbol{\lambda}) = L_0(\mathbf{x}, \boldsymbol{\lambda}) + (\mathsf{A}\mathbf{x} - \mathbf{b} + \mathsf{B}^T\boldsymbol{\lambda})^T\mathbf{d} + \frac{1}{2}\mathbf{d}^T\mathsf{A}\mathbf{d}. \tag{3.22}$$

The Lagrangian function is defined in such a way that if considered as a function of $\mathbf{x}$, then its Hessian and its restriction to Ω_E are exactly those of f, but its gradient $\nabla_{\mathbf{x}} L_0(\mathbf{x}, \boldsymbol{\lambda})$ varies depending on the choice of $\boldsymbol{\lambda}$. It simply follows that if f is convex, then L_0 is convex for any fixed $\boldsymbol{\lambda}$, and the global minimizer of L_0 with respect to $\mathbf{x}$ also varies with $\boldsymbol{\lambda}$. We shall see that it is possible to give conditions on A, B, and $\mathbf{b}$ such that with a suitable choice $\boldsymbol{\lambda} = \widehat{\boldsymbol{\lambda}}$, the solution of the constrained minimization problem (3.17) reduces to the unconstrained minimization of L_0 as in Fig. 3.3.

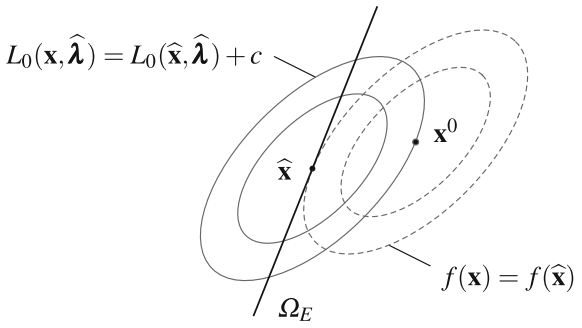

Fig. 3.3 Geometric illustration of the Lagrangian function

3.4.1 Optimality Conditions

The main questions concerning the optimality and solvability conditions of (3.17) are answered by the next proposition.

Proposition 3.6 *Let the equality constrained problem (3.17) be defined by an SPS or SPD matrix* $\mathsf{A} \in \mathbb{R}^{n\times n}$*, a constraint matrix* $\mathsf{B} \in \mathbb{R}^{m\times n}$ *the column rank of which is less than n, and vectors* $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{c} \in \mathrm{Im}\mathsf{B}$*. Then the following statements hold:*
(i) A vector $\overline{\mathbf{x}} \in \Omega_E$ *is a solution of (3.17) if and only if*

$$(\mathsf{A}\overline{\mathbf{x}} - \mathbf{b})^T\mathbf{d} = 0 \tag{3.23}$$

for any $\mathbf{d} \in \mathrm{Ker}\mathsf{B}$.
(ii) A vector $\overline{\mathbf{x}} \in \Omega_E$ *is a solution of (3.17) if and only if there is a vector* $\overline{\boldsymbol{\lambda}} \in \mathbb{R}^m$ *such that*

$$\mathsf{A}\overline{\mathbf{x}} - \mathbf{b} + \mathsf{B}^T\overline{\boldsymbol{\lambda}} = \mathbf{o}. \tag{3.24}$$

Proof (i) Let $\overline{\mathbf{x}}$ be a solution of the equality constrained minimization problem (3.17), so that for any $\mathbf{d} \in \mathrm{Ker}\mathsf{B}$ and $\alpha \in \mathbb{R}$

$$0 \le f(\overline{\mathbf{x}} + \alpha\mathbf{d}) - f(\overline{\mathbf{x}}) = \alpha(\mathsf{A}\overline{\mathbf{x}} - \mathbf{b})^T\mathbf{d} + \frac{\alpha^2}{2}\mathbf{d}^T\mathsf{A}\mathbf{d}. \tag{3.25}$$

For sufficiently small values of α and $(\mathsf{A}\overline{\mathbf{x}} - \mathbf{b})^T\mathbf{d} \ne 0$, the sign of the right-hand side of (3.25) is determined by the sign of $\alpha(\mathsf{A}\overline{\mathbf{x}} - \mathbf{b})^T\mathbf{d}$. Since we can choose the sign of α arbitrarily and the right-hand side of (3.25) is nonnegative, we conclude that (3.23) holds for any $\mathbf{d} \in \mathrm{Ker}\mathsf{B}$.

Let us now assume that (3.23) holds for a vector $\overline{\mathbf{x}} \in \Omega_E$. Then

$$f(\overline{\mathbf{x}} + \mathbf{d}) - f(\overline{\mathbf{x}}) = \frac{1}{2}\mathbf{d}^T\mathsf{A}\mathbf{d} \ge 0$$

for any $\mathbf{d} \in \mathrm{Ker}\mathsf{B}$, so that $\overline{\mathbf{x}}$ is a solution of (3.17).
(ii) Let $\overline{\mathbf{x}}$ be a solution of (3.17), so that by (i) $\overline{\mathbf{x}}$ satisfies (3.23) for any $\mathbf{d} \in \mathrm{Ker}\mathsf{B}$. The latter condition is by (2.30) equivalent to $\mathsf{A}\overline{\mathbf{x}} - \mathbf{b} \in \mathrm{Im}\mathsf{B}^{\mathrm{T}}$, so that there is $\overline{\boldsymbol{\lambda}} \in \mathbb{R}^m$ such that (3.24) holds.

If there are $\overline{\boldsymbol{\lambda}}$ and $\overline{\mathbf{x}} \in \Omega_E$ such that (3.24) holds, then by Taylor's expansion (3.22)

$$f(\overline{\mathbf{x}} + \mathbf{d}) - f(\overline{\mathbf{x}}) = L_0(\overline{\mathbf{x}} + \mathbf{d}, \overline{\boldsymbol{\lambda}}) - L_0(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}) = \frac{1}{2}\mathbf{d}^T\mathsf{A}\mathbf{d} \ge 0$$

for any $\mathbf{d} \in \mathrm{Ker}\mathsf{B}$, so $\overline{\mathbf{x}}$ is a solution of the equality constrained problem (3.17). □

The conditions (ii) of Proposition 3.6 are known as the *Karush–Kuhn–Tucker (KKT) conditions* for the solution of the equality constrained problem (3.17). If $\overline{\mathbf{x}} \in \Omega_E$ and $\overline{\boldsymbol{\lambda}} \in \mathbb{R}^m$ satisfy (3.24), then $(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}})$ is called a *KKT pair* of problem (3.17).

Its second component $\overline{\boldsymbol{\lambda}}$ is called the vector of *Lagrange multipliers* or simply the *multiplier*. We shall often use the notation $\widehat{\mathbf{x}}$ or $\widehat{\boldsymbol{\lambda}}$ to denote the components of a KKT pair that are uniquely determined.

Proposition 3.6 has a simple geometrical interpretation. The condition (3.23) requires that the gradient of f at a solution $\overline{\mathbf{x}}$ is orthogonal to $\mathrm{Ker}\mathsf{B}$, the set of feasible directions of Ω_E, so that there is no feasible decrease direction as illustrated in Fig. 3.4. Since $\mathbf{d}$ is by (2.30) orthogonal to $\mathrm{Ker}\mathsf{B}$ if and only if $\mathbf{d} \in \mathrm{Im}\mathsf{B}^T$, it follows that (3.23) is equivalent to the possibility to choose $\overline{\boldsymbol{\lambda}}$ so that $\nabla_{\mathbf{x}} L_0(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}) = \mathbf{o}$. If f is convex, then the latter condition is equivalent to the condition for the unconstrained minimizer of L_0 with respect to $\mathbf{x}$ as illustrated in Fig. 3.5.

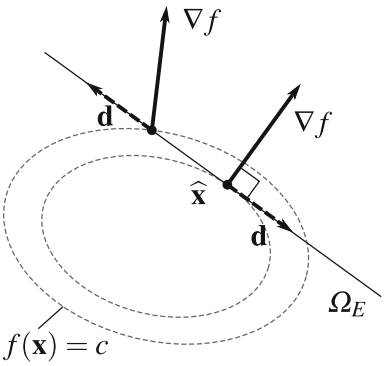

Fig. 3.4 Solvability condition (i)

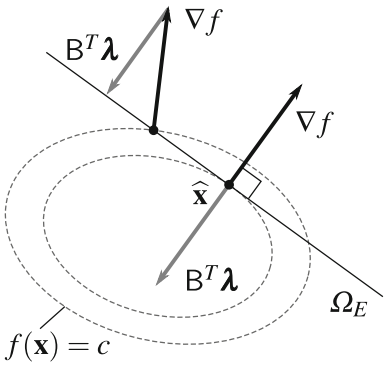

Fig. 3.5 Solvability condition (ii)

Notice that if f is convex, then the vector of Lagrange multipliers which is the component of a KKT pair modifies the linear term of the original problem in such a way that the solution of the unconstrained modified problem is exactly the same as the solution of the original constrained problem. In terms of mechanics, if the original problem describes the equilibrium of a constrained elastic body subject to traction, then the modified problem is unconstrained with the constraints replaced by the reaction forces.

3.4.2 Existence and Uniqueness

Using the optimality conditions of Sect. 3.4.1, we can formulate the conditions that guarantee the existence or uniqueness of a solution of (3.17).

Proposition 3.7 *Let the equality constrained problem (3.17) be defined by an SPS or SPD matrix* $\mathsf{A} \in \mathbb{R}^{n\times n}$*, a constraint matrix* $\mathsf{B} \in \mathbb{R}^{m\times n}$ *the column rank of which is less than n, and vectors* $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{c} \in \mathrm{Im}\mathsf{B}$*. Let* R *denote a matrix the columns of which span* $\mathrm{Ker}\mathsf{A}$*. Then the following statements hold:*

(i) If A *is an SPS matrix, then problem (3.17) has a solution if and only if*

$$\mathsf{R}^T\mathbf{b} \in \mathrm{Im}(\mathsf{R}^T\mathsf{B}^T). \tag{3.26}$$

(ii) If $\mathsf{A}|\mathrm{Ker}\mathsf{B}$ *is positive definite, then problem (3.17) has a unique solution.*
(iii) If $(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}})$ *and* $(\overline{\mathbf{y}}, \overline{\boldsymbol{\mu}})$ *are KKT couples for problem (3.17), then*

$$\overline{\mathbf{x}} - \overline{\mathbf{y}} \in \mathrm{Ker}\mathsf{A} \cap \mathrm{Span}\{\mathbf{b}\}^{\perp} \quad \text{and} \quad \overline{\boldsymbol{\lambda}} - \overline{\boldsymbol{\mu}} \in \mathrm{Ker}\mathsf{B}^T.$$

In particular, if problem (3.17) has a solution and

$$\mathrm{Ker}\mathsf{B}^T = \{\mathbf{o}\},$$

then there is a unique Lagrange multiplier $\widehat{\boldsymbol{\lambda}}$.

Proof (i) Using Proposition 3.6(ii), we have that problem (3.17) has a solution if and only if there is $\boldsymbol{\lambda}$ such that $\mathbf{b} - \mathsf{B}^T\boldsymbol{\lambda} \in \mathrm{Im}\mathsf{A}$, or, equivalently, that $\mathbf{b} - \mathsf{B}^T\boldsymbol{\lambda}$ is orthogonal to $\mathrm{Ker}\mathsf{A}$. The latter condition reads $\mathsf{R}^T\mathbf{b} - \mathsf{R}^T\mathsf{B}^T\boldsymbol{\lambda} = \mathbf{o}$ and can be rewritten as (3.26).
(ii) First observe that if $\mathsf{A}|\mathrm{Ker}\mathsf{B}$ is positive definite, then $f|\mathrm{Ker}\mathsf{B}$ is strictly convex by Proposition 3.2 and it is easy to check that $f|\Omega_E$ is strictly convex. Since Ω_E is closed, convex, and nonempty, it follows by Proposition 3.4(i) that the equality constrained problem (3.17) has a unique solution.
(iii) First observe that $\mathrm{Ker}\mathsf{B} = \{\mathbf{x} - \mathbf{y} : \mathbf{x}, \mathbf{y} \in \Omega_E\}$ and that f is convex on $\mathrm{Ker}\mathsf{B}$ by the assumption and Proposition 3.2. Thus if $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ are any solutions of (3.17), then the left relation follows by Proposition 3.3(ii). The rest follows by a simple analysis of the KKT conditions (3.24). □

If B is not a full row rank matrix and $\overline{\boldsymbol{\lambda}}$ is a Lagrange multiplier for (3.17), then by Proposition 3.7(iii) any Lagrange multiplier $\boldsymbol{\lambda}$ can be expressed in the form

$$\boldsymbol{\lambda} = \overline{\boldsymbol{\lambda}} + \mathbf{d}, \quad \mathbf{d} \in \mathrm{Ker}\mathsf{B}^T. \tag{3.27}$$

The Lagrange multiplier $\boldsymbol{\lambda}_{\mathrm{LS}}$ which minimizes the Euclidean norm is called the *least square Lagrange multiplier*; it is a unique multiplier which belongs to $\mathrm{Im}\mathsf{B}$. If $\overline{\boldsymbol{\lambda}}$ is a vector of Lagrange multipliers, then $\boldsymbol{\lambda}_{\mathrm{LS}}$ can be evaluated by

$$\boldsymbol{\lambda}_{\mathrm{LS}} = \left(\mathsf{B}^{\dagger}\right)^T \mathsf{B}^T\overline{\boldsymbol{\lambda}} \tag{3.28}$$

and

$$\overline{\boldsymbol{\lambda}} = \boldsymbol{\lambda}_{\mathrm{LS}} + \mathbf{d}, \ \mathbf{d} \in \mathrm{Ker}\mathsf{B}^T.$$

If A is positive definite, then the unique solution $\widehat{\mathbf{x}}$ of (3.17) is by Proposition 3.6 fully determined by the matrix equation

$$\begin{bmatrix} \mathsf{A} & \mathsf{B}^T \\ \mathsf{B} & \mathsf{O} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{c} \end{bmatrix}, \tag{3.29}$$

which is known as the *Karush–Kuhn–Tucker system*, briefly *KKT system* or *KKT conditions for the equality constrained problem* (3.17). Proposition 3.6 does not require that the related KKT system is nonsingular, in agreement with observation that the solution of the equality constrained problem should not depend on the description of Ω_E.

3.4.3 Sensitivity

The Lagrange multipliers emerged in Proposition 3.6 as auxiliary variables which nobody had asked for, but which turned out to be useful in alternative formulations of the optimality conditions. However, it turns out that the Lagrange multipliers frequently have an interesting interpretation in specific practical contexts, as we have mentioned at the end of Sect. 3.4.1, where we briefly described their mechanical interpretation. Here we show that if they are uniquely determined by the KKT conditions (3.29), then they are related to the rates of change of the optimal cost due to the violation of constraints.

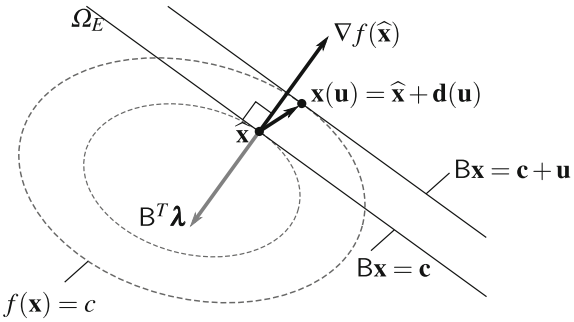

Fig. 3.6 Minimization with perturbed constraints

Let us assume that A and B are positive definite and full rank matrices, respectively, so that there is a unique KKT couple $(\widehat{\mathbf{x}}, \widehat{\boldsymbol{\lambda}})$ of the equality constrained problem (3.17). For $\mathbf{u} \in \mathbb{R}^m$, let us consider also the perturbed problem

$$\min_{\mathsf{B}\mathbf{x}=\mathbf{c}+\mathbf{u}} f(\mathbf{x})$$

as in Fig. 3.6. Its solution $\mathbf{x}(\mathbf{u})$ and the corresponding vector of Lagrange multipliers $\boldsymbol{\lambda}(\mathbf{u})$ are fully determined by the KKT conditions

$$\begin{bmatrix} \mathsf{A} & \mathsf{B}^T \\ \mathsf{B} & \mathsf{O} \end{bmatrix} \begin{bmatrix} \mathbf{x}(\mathbf{u}) \\ \boldsymbol{\lambda}(\mathbf{u}) \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{c}+\mathbf{u} \end{bmatrix},$$

so that

$$\begin{bmatrix} \mathbf{x}(\mathbf{u}) \\ \boldsymbol{\lambda}(\mathbf{u}) \end{bmatrix} = \begin{bmatrix} \mathsf{A} & \mathsf{B}^T \\ \mathsf{B} & \mathsf{O} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{b} \\ \mathbf{c}+\mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathsf{A} & \mathsf{B}^T \\ \mathsf{B} & \mathsf{O} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{b} \\ \mathbf{c} \end{bmatrix} + \begin{bmatrix} \mathsf{A} & \mathsf{B}^T \\ \mathsf{B} & \mathsf{O} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{o} \\ \mathbf{u} \end{bmatrix}.$$

First observe that $\mathbf{d}(\mathbf{u}) = \mathbf{x}(\mathbf{u}) - \widehat{\mathbf{x}}$ satisfies

$$\mathsf{B}\mathbf{d}(\mathbf{u}) = \mathsf{B}\mathbf{x}(\mathbf{u}) - \mathsf{B}\widehat{\mathbf{x}} = \mathbf{u},$$

so that we can use $\nabla f(\widehat{\mathbf{x}}) = -\mathsf{B}^T\widehat{\boldsymbol{\lambda}}$ to approximate the change of optimal cost by

$$\nabla f(\widehat{\mathbf{x}})^T \mathbf{d}(\mathbf{u}) = -(\mathsf{B}^T\widehat{\boldsymbol{\lambda}})^T \mathbf{d}(\mathbf{u}) = -\widehat{\boldsymbol{\lambda}}^T \mathsf{B}\mathbf{d}(\mathbf{u}) = -\widehat{\boldsymbol{\lambda}}^T \mathbf{u}.$$

It follows that $-[\widehat{\boldsymbol{\lambda}}]_i$ can be used to approximate the change of the optimal cost due to the violation of the ith constraint by $[\mathbf{u}]_i$.

To give a more detailed analysis of the sensitivity of the optimal cost with respect to the violation of constraints, let us define for each $\mathbf{u} \in \mathbb{R}^m$ the *primal function*

$$p(\mathbf{u}) = f\left(\mathbf{x}(\mathbf{u})\right).$$

Observing that $\widehat{\mathbf{x}} = \mathbf{x}(\mathbf{o})$ and using the explicit formula (2.4) to evaluate the inverse of the KKT system, we get

$$\mathbf{x}(\mathbf{u}) = \widehat{\mathbf{x}} + \mathsf{A}^{-1}\mathsf{B}^T\mathsf{S}^{-1}\mathbf{u},$$

where $\mathsf{S} = \mathsf{B}\mathsf{A}^{-1}\mathsf{B}^T$ denotes the Schur complement matrix. Thus

$$\mathbf{x}(\mathbf{u}) - \widehat{\mathbf{x}} = \mathsf{A}^{-1}\mathsf{B}^T\mathsf{S}^{-1}\mathbf{u},$$

so that

$$\begin{aligned} p(\mathbf{u}) - p(\mathbf{o}) &= f\left(\mathbf{x}(\mathbf{u})\right) - f(\widehat{\mathbf{x}}) \\ &= \nabla f(\widehat{\mathbf{x}})^T \big(\mathbf{x}(\mathbf{u}) - \widehat{\mathbf{x}}\big) + \frac{1}{2}\big(\mathbf{x}(\mathbf{u}) - \widehat{\mathbf{x}}\big)^T \mathsf{A}\big(\mathbf{x}(\mathbf{u}) - \widehat{\mathbf{x}}\big) \\ &= \nabla f(\widehat{\mathbf{x}})^T \mathsf{A}^{-1}\mathsf{B}^T\mathsf{S}^{-1}\mathbf{u} + \frac{1}{2}\mathbf{u}^T\mathsf{S}^{-1}\mathsf{B}\mathsf{A}^{-1}\mathsf{B}^T\mathsf{S}^{-1}\mathbf{u}. \end{aligned}$$

It follows that the gradient of the primal function p at $\mathbf{o}$ is given by

$$\nabla p(\mathbf{o}) = \big(\nabla f(\widehat{\mathbf{x}})^T \mathsf{A}^{-1}\mathsf{B}^T\mathsf{S}^{-1}\big)^T = \mathsf{S}^{-1}\mathsf{B}\mathsf{A}^{-1}\nabla f(\widehat{\mathbf{x}}).$$

Recalling that $\nabla f(\widehat{\mathbf{x}}) = -\mathsf{B}^T\widehat{\boldsymbol{\lambda}}$, we get

$$\nabla p(\mathbf{o}) = -\mathsf{S}^{-1}\mathsf{B}\mathsf{A}^{-1}\mathsf{B}^T\widehat{\boldsymbol{\lambda}} = -\widehat{\boldsymbol{\lambda}}. \tag{3.30}$$

The analysis shows that the decrease of the total differential of f outside Ω_E near $\widehat{\mathbf{x}}$ is compensated by the increase of $\widehat{\boldsymbol{\lambda}}^T(\mathsf{B}\mathbf{x} - \mathbf{c})$. See also Fig. 3.3. The components of $\widehat{\boldsymbol{\lambda}}$ are also called *shadow prices* after their interpretation in economy.

3.5 Inequality Constrained Problems

Let us now consider the problems

$$\min_{\mathbf{x}\in\Omega_I} f(\mathbf{x}), \quad \Omega_I = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{h}(\mathbf{x}) \le \mathbf{o}\}, \tag{3.31}$$

where f is a quadratic function defined by (3.3) and the constraints are defined by continuously differentiable convex functions $h_i(\mathbf{x}) = [\mathbf{h}(\mathbf{x})]_i$, $i = 1, \ldots, s$, that satisfy $\nabla h_i(\mathbf{x}) \neq \mathbf{o}$ when $h_i(\mathbf{x}) = 0$. In our applications, h_i are either linear forms

$$h_i(\mathbf{x}) = \mathbf{b}_i^T\mathbf{x} - c_i, \quad c_i \in \mathbb{R},$$

or strictly convex separable quadratic functions, i.e.,

$$h_i(\mathbf{x}) = (\mathbf{x}_i - \mathbf{y}_i)^T\mathsf{H}_i(\mathbf{x}_i - \mathbf{y}_i) - c_i, \quad \mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^2, \quad \mathsf{H}_i \ \text{ SPD}, \quad c_i > 0.$$

We assume that Ω_I is nonempty. If the definition of Ω_I includes a quadratic inequality, we call (3.31) the QCQP (Quadratic Constraints Quadratic Cost) problem.

At any feasible point $\mathbf{x}$, we define the *active set*

$$\mathscr{A}(\mathbf{x}) = \{i \in \{1, \ldots, s\} : \ h_i(\mathbf{x}) = 0\}.$$

In particular, if $\overline{\mathbf{x}}$ is a solution of (3.31) with $\mathbf{h}(\mathbf{x}) = \mathsf{B}\mathbf{x} - \mathbf{c}$, $\mathsf{B} \in \mathbb{R}^{s\times n}$, then each feasible direction of $\overline{\Omega}_E = \{\mathbf{x} \in \mathbb{R}^n : \ [\mathsf{B}\mathbf{x}]_{\mathscr{A}(\overline{\mathbf{x}})} = \mathbf{c}_{\mathscr{A}(\overline{\mathbf{x}})}\}$ at $\overline{\mathbf{x}}$ is a feasible direction of Ω_I at $\overline{\mathbf{x}}$. Using the arguments of Sect. 3.4.1, we get that $\overline{\mathbf{x}}$ is also a solution of the equality constrained problem

$$\min_{\mathbf{x}\in\overline{\Omega}_E} f(\mathbf{x}), \quad \overline{\Omega}_E = \{\mathbf{x} \in \mathbb{R}^n : \ [\mathsf{B}\mathbf{x}]_{\mathscr{A}(\overline{\mathbf{x}})} = \mathbf{c}_{\mathscr{A}(\overline{\mathbf{x}})}\}. \tag{3.32}$$

Thus (3.31) is a more difficult problem than the equality constrained problem (3.17) as its solution necessarily enhances the identification of $\mathscr{A}(\overline{\mathbf{x}})$.

3.5.1 Optimality Conditions for Linear Constraints

We shall start our exposition with the following optimality conditions.

Proposition 3.8 *Let the inequality constrained problem (3.31) be defined by an SPS or SPD symmetric matrix* $\mathsf{A} \in \mathbb{R}^{n\times n}$*, the constraint matrix* $\mathsf{B} \in \mathbb{R}^{m\times n}$*, and the vectors* $\mathbf{b}, \mathbf{c} \in \mathbb{R}^n$*. Let* $\Omega_I \neq \emptyset$*. Then the following statements hold:*
(i) $\overline{\mathbf{x}} \in \Omega_I$ *is a solution of (3.31) if and only if*

$$(\mathsf{A}\overline{\mathbf{x}} - \mathbf{b})^T\mathbf{d} \geq 0 \tag{3.33}$$

for any feasible direction $\mathbf{d}$ *of* Ω_I *at* $\overline{\mathbf{x}}$*.*
(ii) $\overline{\mathbf{x}} \in \Omega_I$ *is a solution of (3.31) with linear inequality constraints if and only if there is* $\overline{\boldsymbol{\lambda}} \in \mathbb{R}^m$ *such that*

$$\overline{\boldsymbol{\lambda}} \geq \mathbf{o}, \quad \mathsf{A}\overline{\mathbf{x}} - \mathbf{b} + \mathsf{B}^T\overline{\boldsymbol{\lambda}} = \mathbf{o}, \quad \textit{and} \quad \overline{\boldsymbol{\lambda}}^T(\mathsf{B}\overline{\mathbf{x}} - \mathbf{c}) = 0. \tag{3.34}$$

Proof (i) Let $\overline{\mathbf{x}}$ be a solution of the inequality constrained problem (3.31) and let $\mathbf{d}$ denote a feasible direction of Ω_I at $\overline{\mathbf{x}}$, so that the right-hand side of

$$f(\overline{\mathbf{x}} + \alpha\mathbf{d}) - f(\overline{\mathbf{x}}) = \alpha(\mathsf{A}\overline{\mathbf{x}} - \mathbf{b})^T\mathbf{d} + \frac{\alpha^2}{2}\mathbf{d}^T\mathsf{A}\mathbf{d} \tag{3.35}$$

is nonnegative for all sufficiently small $\alpha > 0$. To prove (3.33), it is enough to take $\alpha > 0$ so small that the nonnegativity of the right-hand side of (3.35) implies that

$$\alpha(\mathsf{A}\overline{\mathbf{x}} - \mathbf{b})^T\mathbf{d} \geq 0.$$

Let us assume that $\overline{\mathbf{x}} \in \Omega_I$ satisfies (3.33) and $\mathbf{x} \in \Omega_I$. Since Ω_I is convex, it follows that $\mathbf{d} = \mathbf{x} - \overline{\mathbf{x}}$ is a feasible direction of Ω_I at $\overline{\mathbf{x}}$, so that, using Taylor's expansion and the assumptions, we have

$$f(\mathbf{x}) - f(\overline{\mathbf{x}}) = (\mathsf{A}\overline{\mathbf{x}} - \mathbf{b})^T\mathbf{d} + \frac{1}{2}\mathbf{d}^T\mathsf{A}\mathbf{d} \geq 0.$$

(ii) Notice any solution $\overline{\mathbf{x}}$ of (3.31) solves (3.32), so that by Proposition 3.6(ii) there is $\mathbf{y}$ such that

$$\mathsf{A}\overline{\mathbf{x}} - \mathbf{b} + \mathsf{B}^T_{\mathscr{A}(\overline{\mathbf{x}})}\mathbf{y} = \mathbf{c}_{\mathscr{A}(\overline{\mathbf{x}})},$$

and $\mathbf{y} \geq \mathbf{o}$ by the arguments based on the sensitivity of the minimum in Sect. 3.4.3. To finish the proof, it is enough to define $\boldsymbol{\lambda}$ as $\mathbf{y}$ padded with zeros.

If (3.34) holds and $\overline{\mathbf{x}} + \mathbf{d} \in \Omega_I$, then $\mathsf{A}\overline{\mathbf{x}} - \mathbf{b} = -\mathsf{B}^T\boldsymbol{\lambda}$ and

$$f(\overline{\mathbf{x}} + \mathbf{d}) - f(\overline{\mathbf{x}}) = (\mathsf{A}\overline{\mathbf{x}} - \mathbf{b})^T\mathbf{d} + \frac{1}{2}\mathbf{d}^T\mathsf{A}\mathbf{d} \geq -\boldsymbol{\lambda}^T\mathsf{B}\mathbf{d} = -\boldsymbol{\lambda}^T(\mathsf{B}(\overline{\mathbf{x}} + \mathbf{d}) - \mathbf{c}) \geq \mathbf{o}. \qquad \square$$

Remark 3.2 Condition (3.33) can be written as a variational inequality

$$(\mathsf{A}\overline{\mathbf{x}})^T(\mathbf{x}-\overline{\mathbf{x}}) \geq \mathbf{b}^T(\mathbf{x}-\overline{\mathbf{x}}), \quad \mathbf{x} \in \Omega_I.$$

The conditions (3.34) are called the *KKT conditions for inequality constraints*. The last of these conditions is called the *condition of complementarity*.

3.5.2 Optimality Conditions for Bound Constrained Problems

A special case of problem (3.31) is the bound constrained problem

$$\min_{\mathbf{x}\in\Omega_B} f(\mathbf{x}), \quad \Omega_B = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq \ell\}, \tag{3.36}$$

where f is a quadratic function defined by (3.3) and $\ell \in \mathbb{R}^n$. The optimality conditions for convex bound constrained problems can be written in a more convenient form.

Proposition 3.9 *Let f be a convex quadratic function defined by (3.3) with a positive semidefinite Hessian A. Then $\overline{\mathbf{x}} \in \Omega_B$ solves (3.36) if and only if*

$$\mathsf{A}\overline{\mathbf{x}} - \mathbf{b} \geq \mathbf{o} \quad \textit{and} \quad (\mathsf{A}\overline{\mathbf{x}} - \mathbf{b})^T(\overline{\mathbf{x}} - \ell) = 0. \tag{3.37}$$

Proof First observe that denoting $\mathsf{B} = -\mathsf{I}_n$, $\mathbf{c} = -\ell$, and

$$\Omega_I = \{\mathbf{x} \in \mathbb{R}^n : \mathsf{B}\mathbf{x} \leq \mathbf{c}\},$$

the bound constrained problem (3.36) becomes the standard inequality constrained problem (3.31) with $\Omega_I = \Omega_B$. Using Proposition 3.11, it follows that $\overline{\mathbf{x}} \in \Omega_B$ is the solution of (3.36) if and only if there is $\boldsymbol{\lambda} \in \mathbb{R}^n$ such that

$$\boldsymbol{\lambda} \geq \mathbf{o}, \quad \mathsf{A}\overline{\mathbf{x}} - \mathbf{b} - \mathsf{I}\boldsymbol{\lambda} = \mathbf{o}, \quad \text{and} \quad \boldsymbol{\lambda}^T(\overline{\mathbf{x}} - \ell) = 0. \tag{3.38}$$

We complete the proof by observing that (3.37) can be obtained from (3.38) and vice versa by substituting $\boldsymbol{\lambda} = \mathsf{A}\overline{\mathbf{x}} - \mathbf{b}$. □

In the proof, we have shown that $\boldsymbol{\lambda} = \nabla f(\overline{\mathbf{x}})$ is a vector of Lagrange multipliers for the constraints $-\mathbf{x} \leq -\ell$, or, equivalently, for $\mathbf{x} \geq \ell$. Notice that the conditions (3.37) require that none of the vectors $\mathbf{s}_i$ is a feasible decrease direction of Ω_B at $\overline{\mathbf{x}}$, where $\mathbf{s}_i$ denotes a vector of the standard basis of $\mathbb{R}^n$ formed by the columns of I_n, $i \in \mathscr{A}(\overline{\mathbf{x}})$.

3.5.3 Optimality Conditions for More General Constraints

If the constraints h_i that define (3.31) are nonlinear, it can happen that their first-order representation by means of the gradients ∇h_i is not adequate. The reason is illustrated in Fig. 3.7, where the *linear cone* $\mathscr{L}_{\Omega_I}(\overline{\mathbf{x}})$ of Ω_I at $\overline{\mathbf{x}}$ on the boundary of Ω_I defined by

$$\mathscr{L}_{\Omega_I}(\overline{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^n : \mathbf{d}^T \nabla h_i(\overline{\mathbf{x}}) \leq 0,\ i \in \mathscr{A}(\overline{\mathbf{x}})\}$$

comprises the whole line, while the *tangent cone* $\mathscr{T}_{\Omega_I}(\overline{\mathbf{x}})$ of Ω_I at $\overline{\mathbf{x}}$ on the boundary of Ω_I defined by

$$\mathscr{T}_{\Omega_I}(\overline{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^n :\ \mathbf{d} = \lim_{i\to\infty} \mathbf{d}^i,\ \overline{\mathbf{x}} + \alpha_i \mathbf{d}^i \in \Omega_I,\ \lim_{i\to\infty} \alpha_i = 0,\ \alpha_i > 0\}$$

comprises only one point. To avoid such pathological situations, we shall assume that $\mathscr{L}_{\Omega_I}(\overline{\mathbf{x}}) = \mathscr{T}_{\Omega_I}(\overline{\mathbf{x}})$. This is also called the *Abadie constraint qualification (ACQ)* [5]. Notice that linear constraints satisfy ACQ. The ACQ assumption reduces the analysis of the conditions of minima to the linear case, so that we can formulate the following proposition.

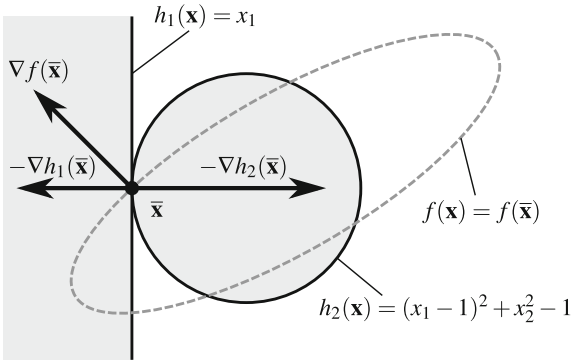

Fig. 3.7 Example of $\Omega_I = \{\overline{\mathbf{x}}\}$, $\mathscr{L}_{\Omega_I}(\overline{\mathbf{x}}) \neq \mathscr{T}_{\Omega_I}(\overline{\mathbf{x}})$

Proposition 3.10 *Let the inequality constrained problem (3.31) be defined by an SPS or SPD matrix* $\mathsf{A} \in \mathbb{R}^{n\times n}$ *and convex differentiable functions* h_i *and let* Ω_I *satisfies ACQ. Then* $\overline{\mathbf{x}} \in \Omega_I$ *is a solution of (3.31) if and only if there is* $\overline{\boldsymbol{\lambda}} \in \mathbb{R}^m$ *such that*

$$\overline{\boldsymbol{\lambda}} \geq \mathbf{o}, \quad \mathsf{A}\overline{\mathbf{x}} - \mathbf{b} + \nabla \mathbf{h}(\overline{\mathbf{x}})\overline{\boldsymbol{\lambda}} = \mathbf{o}, \quad \textit{and} \quad \overline{\boldsymbol{\lambda}}^T \mathbf{h}(\overline{\mathbf{x}}) = 0. \tag{3.39}$$

Proof First notice that due to the definition of tangential cone, $\overline{\mathbf{x}}$ solves (3.31) if and only if

$$f(\overline{\mathbf{x}}) = \min_{\overline{\mathbf{x}}+\mathscr{T}_{\Omega_I}(\overline{\mathbf{x}})} f(\mathbf{x}).$$

Since we assume $\mathscr{T}_{\Omega_I}(\overline{\mathbf{x}}) = \mathscr{L}_{\Omega_I}(\overline{\mathbf{x}})$, the letter problem has the same solution as

$$\min_{\overline{\mathbf{x}}+\mathscr{L}_{\Omega_I}(\overline{\mathbf{x}})} f(\mathbf{x}).$$

Using Proposition 3.8, we get that $\overline{\mathbf{x}} \in \Omega_I$ solves (3.31) if and only if $\overline{\mathbf{x}}$ satisfies (3.39). □

3.5.4 Existence and Uniqueness

In our discussion of the existence and uniqueness results for the inequality constrained QP problem (3.31), we restrict our attention to the following results that are useful in our applications.

Proposition 3.11 *Let the inequality constrained problem (3.31) be defined by convex functions h_i and f. Let $\mathscr{C}$ denote the cone of recession directions of the nonempty feasible set Ω_I. Then the following statements hold:*
(i) If problem (3.31) has a solution, then

$$\mathbf{d}^T\mathbf{b} \le 0 \quad for \quad \mathbf{d} \in \mathscr{C} \cap \mathrm{Ker}\mathsf{A}. \tag{3.40}$$

(ii) If the constraints are linear, then (3.40) is sufficient for the existence of minima.
(iii) If the constraints are linear and $(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}})$ and $(\overline{\mathbf{y}}, \overline{\boldsymbol{\mu}})$ are KKT couples for (3.31), then

$$\overline{\mathbf{x}} - \overline{\mathbf{y}} \in \mathrm{Ker}\mathsf{A} \cap \mathrm{Span}\{\mathbf{b}\}^{\perp} \quad and \quad \overline{\boldsymbol{\lambda}} - \overline{\boldsymbol{\mu}} \in \mathrm{Ker}\mathsf{B}^T. \tag{3.41}$$

(iv) If A is positive definite, then the inequality constrained minimization problem (3.31) has a unique solution.

Proof (i) Let $\overline{\mathbf{x}}$ be a global solution of the inequality constrained minimization problem (3.31), and recall that

$$f(\overline{\mathbf{x}} + \alpha\mathbf{d}) - f(\overline{\mathbf{x}}) = \alpha(\mathsf{A}\overline{\mathbf{x}} - \mathbf{b})^T\mathbf{d} + \frac{\alpha^2}{2}\mathbf{d}^T\mathsf{A}\mathbf{d} \tag{3.42}$$

for any $\mathbf{d} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$. Taking $\mathbf{d} \in \mathscr{C} \cap \mathrm{Ker}\mathsf{A}$, (3.42) reduces to

$$f(\overline{\mathbf{x}} + \alpha\mathbf{d}) - f(\overline{\mathbf{x}}) = -\alpha\mathbf{b}^T\mathbf{d},$$

which is nonnegative for any $\alpha \ge 0$ if and only if $\mathbf{b}^T\mathbf{d} \le 0$.
(ii) See Dostál [7].

(iii) The first inclusion of (3.41) holds by Proposition 3.3(ii) for the solutions of any convex problem. The inclusion for multipliers follows by the KKT condition (3.34). (iv) If A is SPD, then f is strictly convex by Proposition 3.2, so by Proposition 3.4 there is a unique minimizer of f subject to $\mathbf{x} \in \Omega_I$. □

3.6 Equality and Inequality Constrained Problems

In the previous sections, we have obtained the results concerning optimization problems with either equality or inequality constraints. Here we extend these results to the optimization problems with both equality and inequality constraints. More formally, we look for

$$\min_{\mathbf{x}\in\Omega_{IE}} f(\mathbf{x}), \quad \Omega_{IE} = \{\mathbf{x} \in \mathbb{R}^n : \ [\mathbf{h}(\mathbf{x})]_{\mathscr{I}} \le \mathbf{o}_{\mathscr{I}}, \ [\mathbf{h}(\mathbf{x})]_{\mathscr{E}} = \mathbf{o}_{\mathscr{E}}\}, \tag{3.43}$$

where f is a quadratic function with the SPS Hessian $\mathsf{A} \in \mathbb{R}^{n\times n}$ and the linear term defined by $\mathbf{b} \in \mathbb{R}^n$, $\mathscr{I}$, $\mathscr{E}$ are the disjunct sets of indices which decompose $\{1, \ldots, m\}$, and the equality and inequality constraints are defined respectively by linear and continuously differentiable convex functions $[\mathbf{h}(\mathbf{x})]_i = h_i(\mathbf{x}), i = 1, \ldots, m$. We assume that $\nabla h_i(\mathbf{x}) \neq \mathbf{o}$ when $h_i(\mathbf{x}) = 0$ and that $\Omega_i \neq \emptyset$. We are especially interested in linear equality constraints and the inequality constraints defined either by linear forms or by strictly convex separable quadratic functions.

If we describe the conditions that define Ω_{IE} in components, we get

$$\Omega_{IE} = \{\mathbf{x} \in \mathbb{R}^n : \ h_i(\mathbf{x}) \le 0, \ i \in \mathscr{I}, \ \mathbf{b}_i^T\mathbf{x} = c_i, \ i \in \mathscr{E}\},$$

which makes sense even for $\mathscr{I} = \emptyset$ or $\mathscr{E} = \emptyset$; we consider the conditions which concern the empty set as always satisfied. For example, $\mathscr{E} = \emptyset$ gives

$$\Omega_{IE} = \{\mathbf{x} \in \mathbb{R}^n : \ \mathbf{h}_i(\mathbf{x}) \le 0, \ i \in \mathscr{I}\},$$

and the kernel of an "empty" matrix is defined by

$$\mathrm{Ker}\mathsf{B}_{\mathscr{E}*} = \{\mathbf{x} \in \mathbb{R}^n : \ \mathbf{b}_i^T\mathbf{x} = 0, \ i \in \mathscr{E}\} = \mathbb{R}^n.$$

If all constraints are linear, then Ω_I is defined by $\mathsf{B} \in \mathbb{R}^{m\times n}$ and $\mathbf{c} \in \mathbb{R}^m$,

$$\mathsf{B} = \begin{bmatrix} \mathbf{b}_1^T \\ \ldots \\ \mathbf{b}_m^T \end{bmatrix} = \begin{bmatrix} \mathsf{B}_I \\ \mathsf{B}_E \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} \mathbf{c}_I \\ \mathbf{c}_E \end{bmatrix},$$

and we get a QP variant of (3.43)

$$\min_{\mathbf{x}\in\Omega_{IE}} f(\mathbf{x}), \quad \Omega_{IE} = \{\mathbf{x} \in \mathbb{R}^n : \ \mathsf{B}_I\mathbf{x} \le \mathbf{c}_I, \ \ \mathsf{B}_E\mathbf{x} = \mathbf{c}_E\}. \tag{3.44}$$

3.6.1 Optimality Conditions

First observe that any equality constraint $\mathbf{b}_i^T\mathbf{x} = c_i,\ i \in \mathscr{E}$ can be replaced by the couple of inequalities $\mathbf{b}_i^T\mathbf{x} \le c_i$ and $-\mathbf{b}_i^T\mathbf{x} \le -c_i$. We can thus use our results obtained by the analysis of the inequality constrained problems in Sect. 3.5 to get similar results for general bound and equality constrained QP problem (3.43).

Proposition 3.12 *Let the quadratic function f be defined by the SPS matrix A and $\mathbf{b} \in \mathbb{R}^n$. Let $\Omega_{IE} \neq \emptyset$ be defined by*

$$\mathbf{h}_{\mathscr{E}}(\mathbf{x}) = \mathsf{B}_E\mathbf{x} - \mathbf{c}_E$$

and convex differential functions $h_i,\ i \in \mathscr{I}$. Then the following statements hold:
(i) $\overline{\mathbf{x}} \in \Omega_{IE}$ is a solution of (3.43) if and only if

$$\nabla f(\overline{\mathbf{x}}) = (\mathsf{A}\overline{\mathbf{x}} - \mathbf{b})^T\mathbf{d} \ge 0 \tag{3.45}$$

for any feasible direction $\mathbf{d}$ of Ω_{IE} at $\overline{\mathbf{x}}$.
(ii) If Ω_{IE} is defined by linear constraints (3.44), then $\overline{\mathbf{x}} \in \Omega_{IE}$ is a solution of (3.43) if and only if there is a vector $\overline{\boldsymbol{\lambda}} \in \mathbb{R}^m$ such that

$$\overline{\boldsymbol{\lambda}}_{\mathscr{I}} \ge \mathbf{o}, \quad \mathsf{A}\overline{\mathbf{x}} - \mathbf{b} + \mathsf{B}^T\overline{\boldsymbol{\lambda}} = \mathbf{o}, \quad \text{and} \quad \overline{\boldsymbol{\lambda}}_{\mathscr{I}}^T[\mathsf{B}\overline{\mathbf{x}} - \mathbf{c}]_{\mathscr{I}} = 0. \tag{3.46}$$

(iii) If Ω_{IE} is a feasible set for the problem (3.43) and the constraints satisfy ACQ, then $\overline{\mathbf{x}} \in \Omega_{IE}$ is a solution of (3.43) if and only if there is a vector $\overline{\boldsymbol{\lambda}} \in \mathbb{R}^m$ such that

$$\overline{\boldsymbol{\lambda}}_{\mathscr{I}} \ge \mathbf{o}, \quad \mathsf{A}\overline{\mathbf{x}} - \mathbf{b} + \nabla\mathbf{h}(\overline{\mathbf{x}})\overline{\boldsymbol{\lambda}} = \mathbf{o}, \quad \text{and} \quad \overline{\boldsymbol{\lambda}}_{\mathscr{I}}^T\mathbf{h}_{\mathscr{I}}(\overline{\mathbf{x}}) = 0. \tag{3.47}$$

Proof First observe that if $\mathscr{E} = \emptyset$, then the statements of the above proposition reduce to Propositions 3.8 and 3.10, and if $\mathscr{I} = \emptyset$, then they reduce to Proposition 3.6. Thus we can assume in the rest of the proof that $\mathscr{I} \neq \emptyset$ and $\mathscr{E} \neq \emptyset$.

As mentioned above, (3.43) may be rewritten also as

$$\min_{\mathbf{x}\in\Omega_I} f(\mathbf{x}), \quad \Omega_I = \{\mathbf{x} \in \mathbb{R}^n :\ [\mathsf{B}\mathbf{x}]_{\mathscr{I}} \le \mathbf{c}_{\mathscr{I}},\ [\mathsf{B}\mathbf{x}]_{\mathscr{E}} \le \mathbf{c}_{\mathscr{E}}, -[\mathsf{B}\mathbf{x}]_{\mathscr{E}} \le -\mathbf{c}_{\mathscr{E}}\}, \tag{3.48}$$

where $\Omega_I = \Omega_{IE}$. Thus the statement (i) is a special case of Proposition 3.8. Observing that we can always ignore one of the inequality constraints related tho the same equality constraint, we get easily (ii) and (iii) from Propositions 3.8 and 3.10, respectively. □

Remark 3.3 Condition (3.45) can be written as a variational inequality

$$(\mathsf{A}\overline{\mathbf{x}})^T(\mathbf{x} - \overline{\mathbf{x}}) \ge \mathbf{b}^T(\mathbf{x} - \overline{\mathbf{x}}), \quad \mathbf{x} \in \Omega_{IE}.$$

3.7 Duality for Quadratic Programming Problems

The duality associates each problem (3.43), which we shall also call *primal problem*, with a maximization problem in Lagrange multipliers that we shall call the *dual problem*. The solution of the dual problem is a Lagrange multiplier of the solution of the primal problem, so that having the solution of the dual problem, we can get the solution of the primal problem by solving an unconstrained problem. Here we limit our attention to the QP problems (3.44), postponing the discussion of more general cases to specific applications.

The cost function of the dual problem is the *dual function*

$$\Theta(\boldsymbol{\lambda}) = \inf_{\mathbf{x}\in\mathbb{R}^n} L_0(\mathbf{x}, \boldsymbol{\lambda}). \tag{3.49}$$

If A is SPD, then L_0 is a quadratic function with the SPD Hessian and Θ is a quadratic function in $\boldsymbol{\lambda}$ which can be defined by an explicit formula. However, in our applications, it often happens that A is only SPS, so the cost function f need not be bounded from below and $-\infty$ can be in the range of the dual function Θ. We resolve this problem by keeping Θ quadratic at the cost of introducing equality constraints.

Proposition 3.13 *Let matrices* A, B, *vectors* $\mathbf{b}$, $\mathbf{c}$, *and index sets* $\mathscr{I}$, $\mathscr{E}$ *be those of the definition of problem (3.44) with* A *positive semidefinite and* $\Omega_{IE} \neq \emptyset$. *Let* $\mathsf{R} \in \mathbb{R}^{n\times d}$ *be a full rank matrix such that*

$$\mathrm{Im}\mathsf{R} = \mathrm{Ker}\mathsf{A},$$

let A^+ *denote an SPS generalized inverse of* A, *and let*

$$\Theta(\boldsymbol{\lambda}) = -\frac{1}{2}\boldsymbol{\lambda}^T\mathsf{B}\mathsf{A}^+\mathsf{B}^T\boldsymbol{\lambda} + \boldsymbol{\lambda}^T(\mathsf{B}\mathsf{A}^+\mathbf{b} - \mathbf{c}) - \frac{1}{2}\mathbf{b}^T\mathsf{A}^+\mathbf{b}. \tag{3.50}$$

Then the following statements hold:
(i) If $(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}})$ *is a KKT pair for (3.44), then* $\overline{\boldsymbol{\lambda}}$ *is a solution of*

$$\max_{\boldsymbol{\lambda}\in\Omega_{BE}} \Theta(\boldsymbol{\lambda}), \quad \Omega_{BE} = \{\boldsymbol{\lambda} \in \mathbb{R}^m : \boldsymbol{\lambda}_{\mathscr{I}} \geq \mathbf{o},\ \mathsf{R}^T\mathsf{B}^T\boldsymbol{\lambda} = \mathsf{R}^T\mathbf{b}\}. \tag{3.51}$$

Moreover, there is $\overline{\boldsymbol{\alpha}} \in \mathbb{R}^d$ *such that* $(\overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\alpha}})$ *is a KKT pair for problem (3.51) and*

$$\overline{\mathbf{x}} = \mathsf{A}^+(\mathbf{b} - \mathsf{B}^T\overline{\boldsymbol{\lambda}}) + \mathsf{R}\overline{\boldsymbol{\alpha}}. \tag{3.52}$$

(ii) If $(\overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\alpha}})$ *is a KKT pair for problem (3.51), then* $\overline{\mathbf{x}}$ *defined by (3.52) is a solution of the equality and inequality constrained problem (3.44).*
(iii) If $(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}})$ *is a KKT pair for problem (3.44), then*

$$f(\overline{\mathbf{x}}) = \Theta(\overline{\boldsymbol{\lambda}}). \tag{3.53}$$

Proof (i) Assume that $(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}})$ is a KKT pair for (3.44), so that $(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}})$ is by Proposition 3.12 a solution of

$$\begin{aligned}
\boldsymbol{\lambda}_{\mathscr{I}} &\geq \mathbf{o}, && (3.54)\\
\nabla_{\mathbf{x}} L_0(\mathbf{x}, \boldsymbol{\lambda}) &= \mathsf{A}\mathbf{x} - \mathbf{b} + \mathsf{B}^T\boldsymbol{\lambda} = \mathbf{o}, && (3.55)\\
[\nabla_{\boldsymbol{\lambda}} L_0(\mathbf{x}, \boldsymbol{\lambda})]_{\mathscr{I}} &= \big[\mathsf{B}\mathbf{x} - \mathbf{c}\big]_{\mathscr{I}} \leq \mathbf{o}, && (3.56)\\
[\nabla_{\boldsymbol{\lambda}} L_0(\mathbf{x}, \boldsymbol{\lambda})]_{\mathscr{E}} &= \big[\mathsf{B}\mathbf{x} - \mathbf{c}\big]_{\mathscr{E}} = \mathbf{o}, && (3.57)\\
\boldsymbol{\lambda}^T_{\mathscr{I}}[\mathsf{B}\mathbf{x} - \mathbf{c}]_{\mathscr{I}} &= 0. && (3.58)
\end{aligned}$$

Notice that given a vector $\boldsymbol{\lambda} \in \mathbb{R}^m$, we can express the condition

$$\mathbf{b} - \mathsf{B}^T\boldsymbol{\lambda} \in \mathrm{Im}\mathsf{A},$$

which guarantees solvability of (3.55) with respect to $\mathbf{x}$, conveniently as

$$\mathsf{R}^T(\mathsf{B}^T\boldsymbol{\lambda} - \mathbf{b}) = \mathbf{o}. \tag{3.59}$$

If the latter condition is satisfied, then we can use any symmetric left generalized inverse A^+ to find all solutions of (3.55) with respect to $\mathbf{x}$ in the form

$$\mathbf{x}(\boldsymbol{\lambda}, \boldsymbol{\alpha}) = \mathsf{A}^+(\mathbf{b} - \mathsf{B}^T\boldsymbol{\lambda}) + \mathsf{R}\boldsymbol{\alpha},\ \boldsymbol{\alpha} \in \mathbb{R}^d,$$

where d is the dimension of $\mathrm{Ker}\mathsf{A}$. After substituting for $\mathbf{x}$ into (3.56)–(3.58), we get

$$\begin{aligned}
[\ -\mathsf{B}\mathsf{A}^+\mathsf{B}^T\boldsymbol{\lambda} + (\mathsf{B}\mathsf{A}^+\mathbf{b} - \mathbf{c}) + \mathsf{B}\mathsf{R}\boldsymbol{\alpha}]_{\mathscr{I}} &\leq \mathbf{o}, && (3.60)\\
[\ -\mathsf{B}\mathsf{A}^+\mathsf{B}^T\boldsymbol{\lambda} + (\mathsf{B}\mathsf{A}^+\mathbf{b} - \mathbf{c}) + \mathsf{B}\mathsf{R}\boldsymbol{\alpha}]_{\mathscr{E}} &= \mathbf{o}, && (3.61)\\
\boldsymbol{\lambda}^T_{\mathscr{I}}\ [\ -\mathsf{B}\mathsf{A}^+\mathsf{B}^T\boldsymbol{\lambda} + (\mathsf{B}\mathsf{A}^+\mathbf{b} - \mathbf{c}) + \mathsf{B}\mathsf{R}\boldsymbol{\alpha}]_{\mathscr{I}} &= 0. && (3.62)
\end{aligned}$$

The formulae in (3.60)–(3.62) look like something that we have already seen. Indeed, introducing the vector of Lagrange multipliers $\boldsymbol{\alpha}$ for (3.59) and denoting

$$\begin{aligned}
\Lambda(\boldsymbol{\lambda}, \boldsymbol{\alpha}) &= \Theta(\boldsymbol{\lambda}) + \boldsymbol{\alpha}^T(\mathsf{R}^T\mathsf{B}^T\boldsymbol{\lambda} - \mathsf{R}^T\mathbf{b})\\
&= -\frac{1}{2}\boldsymbol{\lambda}^T\mathsf{B}\mathsf{A}^+\mathsf{B}^T\boldsymbol{\lambda} + \boldsymbol{\lambda}^T(\mathsf{B}\mathsf{A}^+\mathbf{b} - \mathbf{c}) - \frac{1}{2}\mathbf{b}^T\mathsf{A}^+\mathbf{b}\\
&\quad + \boldsymbol{\alpha}^T(\mathsf{R}^T\mathsf{B}^T\boldsymbol{\lambda} - \mathsf{R}^T\mathbf{b}),
\end{aligned}$$

$$\mathbf{g} = \nabla_{\boldsymbol{\lambda}}\Lambda(\boldsymbol{\lambda}, \boldsymbol{\alpha}) = -\mathsf{B}\mathsf{A}^+\mathsf{B}^T\boldsymbol{\lambda} + (\mathsf{B}\mathsf{A}^+\mathbf{b} - \mathbf{c}) + \mathsf{B}\mathsf{R}\boldsymbol{\alpha},$$

we can rewrite the relations (3.60)–(3.62) as

$$\mathbf{g}_{\mathscr{I}} \leq \mathbf{o}, \quad \mathbf{g}_{\mathscr{E}} = \mathbf{o}, \quad \text{and} \quad \boldsymbol{\lambda}^T_{\mathscr{I}}\mathbf{g}_{\mathscr{I}} = 0. \tag{3.63}$$

Comparing (3.63) with the KKT conditions for the bound and equality constrained problem, we conclude that (3.63) are the KKT conditions for

$$\max \Theta(\boldsymbol{\lambda}) \quad \text{subject to} \quad \mathsf{R}^T\mathsf{B}^T\boldsymbol{\lambda} - \mathsf{R}^T\mathbf{b} = \mathbf{o} \quad \text{and} \quad \boldsymbol{\lambda}_{\mathscr{I}} \geq \mathbf{o}. \tag{3.64}$$

We have thus proved that if $(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}})$ solves (3.54)–(3.58), then $\overline{\boldsymbol{\lambda}}$ is a feasible vector for problem (3.64) which satisfies the related KKT conditions. Recalling that A^+ is by the assumption symmetric positive semidefinite, so that $\mathsf{B}\mathsf{A}^+\mathsf{B}^T$ is also positive semidefinite, we conclude that $\overline{\boldsymbol{\lambda}}$ solves (3.51). Moreover, we have shown that any solution $\overline{\mathbf{x}}$ can be obtained in the form (3.52) with a KKT pair $(\overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\alpha}})$, where $\overline{\boldsymbol{\alpha}}$ is a vector of the Lagrange multipliers for the equality constraints in (3.51).
(ii) Let $(\overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\alpha}})$ be a KKT pair for problem (3.51), so that $(\overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\alpha}})$ satisfies (3.59)–(3.62) and $\overline{\boldsymbol{\lambda}}_{\mathscr{I}} \geq \mathbf{o}$. If we denote

$$\overline{\mathbf{x}} = \mathsf{A}^+(\mathbf{b} - \mathsf{B}^T\overline{\boldsymbol{\lambda}}) + \mathsf{R}\overline{\boldsymbol{\alpha}},$$

we can use (3.60)–(3.62) to verify directly that $\overline{\mathbf{x}}$ is feasible and $(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}})$ satisfies the complementarity conditions, respectively. Finally, using (3.59), we get that there is $\mathbf{y} \in \mathbb{R}^n$ such that

$$\mathbf{b} - \mathsf{B}^T\overline{\boldsymbol{\lambda}} = \mathsf{A}\mathbf{y}.$$

Thus

$$\begin{aligned}\mathsf{A}\overline{\mathbf{x}} - \mathbf{b} + \mathsf{B}^T\overline{\boldsymbol{\lambda}} &= \mathsf{A}\big(\mathsf{A}^+(\mathbf{b} - \mathsf{B}^T\overline{\boldsymbol{\lambda}}) + \mathsf{R}\overline{\boldsymbol{\alpha}}\big) - \mathbf{b} + \mathsf{B}^T\overline{\boldsymbol{\lambda}} \\ &= \mathsf{A}\mathsf{A}^+\mathsf{A}\mathbf{y} - \mathbf{b} + \mathsf{B}^T\overline{\boldsymbol{\lambda}} = \mathbf{b} - \mathsf{B}^T\overline{\boldsymbol{\lambda}} - \mathbf{b} + \mathsf{B}^T\overline{\boldsymbol{\lambda}} = \mathbf{o},\end{aligned}$$

which proves that $(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}})$ is a KKT pair for (3.44).
(iii) Let $(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}})$ be a KKT pair for (3.44). Using the feasibility condition (3.57) and the complementarity condition (3.58), we get

$$\overline{\boldsymbol{\lambda}}^T(\mathsf{B}\overline{\mathbf{x}} - \mathbf{c}) = \overline{\boldsymbol{\lambda}}_{\mathscr{E}}^T[\mathsf{B}\overline{\mathbf{x}} - \mathbf{c}]_{\mathscr{E}} + \overline{\boldsymbol{\lambda}}_{\mathscr{I}}^T[\mathsf{B}\overline{\mathbf{x}} - \mathbf{c}]_{\mathscr{I}} = 0.$$

Hence

$$f(\overline{\mathbf{x}}) = f(\overline{\mathbf{x}}) + \overline{\boldsymbol{\lambda}}^T(\mathsf{B}\overline{\mathbf{x}} - \mathbf{c}) = L_0(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}).$$

Next recall that if $(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}})$ is a KKT pair, then

$$\nabla_{\mathbf{x}} L_0(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}) = \mathbf{o}.$$

Since L_0 is convex, the latter is the gradient condition for the unconstrained minimizer of L_0 with respect to $\mathbf{x}$; therefore

$$L_0(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}) = \min_{\mathbf{x} \in \mathbb{R}^n} L_0(\mathbf{x}, \overline{\boldsymbol{\lambda}}) = \Theta(\overline{\boldsymbol{\lambda}}).$$

Thus

$$f(\overline{\mathbf{x}}) = L_0(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}) = \Theta(\overline{\boldsymbol{\lambda}}).$$

□

Since the constant term is not essential in our applications and we formulate our algorithms for minimization problems, we shall consider the function

$$\theta(\boldsymbol{\lambda}) = -\Theta(\boldsymbol{\lambda}) - \frac{1}{2}\mathbf{b}^T\mathsf{A}^+\mathbf{b} = \frac{1}{2}\boldsymbol{\lambda}^T\mathsf{B}\mathsf{A}^+\mathsf{B}^T\boldsymbol{\lambda} - \boldsymbol{\lambda}^T(\mathsf{B}\mathsf{A}^+\mathbf{b} - \mathbf{c}), \tag{3.65}$$

so that

$$\arg\min_{\boldsymbol{\lambda}\in\Omega_{BE}} \theta(\boldsymbol{\lambda}) = \arg\max_{\boldsymbol{\lambda}\in\Omega_{BE}} \Theta(\boldsymbol{\lambda}).$$

3.7.1 Uniqueness of a KKT Pair

We shall complete our exposition of duality by formulating the results concerning the uniqueness of the solution for the *constrained dual problem*

$$\min_{\boldsymbol{\lambda}\in\Omega_{BE}} \theta(\boldsymbol{\lambda}), \quad \Omega_{BE} = \{\boldsymbol{\lambda} \in \mathbb{R}^m : \boldsymbol{\lambda}_{\mathscr{I}} \geq \mathbf{o},\ \mathsf{R}^T\mathsf{B}^T\boldsymbol{\lambda} = \mathsf{R}^T\mathbf{b}\}, \tag{3.66}$$

where θ is defined by (3.65).

Proposition 3.14 *Let the matrices* A, B, *the vectors* $\mathbf{b}$, $\mathbf{c}$, *and the index sets* $\mathscr{I}$, $\mathscr{E}$ *be those from the definition of problem (3.44) with* A *positive semidefinite,* $\Omega_{IE} \neq \emptyset$, *and* $\Omega_{BE} \neq \emptyset$. *Let* $\mathsf{R} \in \mathbb{R}^{n\times d}$ *be a full rank matrix such that*

$$\mathrm{Im}\mathsf{R} = \mathrm{Ker}\mathsf{A}.$$

Then the following statements hold:
(i) If B^T *and* $\mathsf{B}\mathsf{R}$ *are full column rank matrices, then there is a unique solution* $\widehat{\boldsymbol{\lambda}}$ *of problem (3.66).*
(ii) If $\widehat{\boldsymbol{\lambda}}$ *is a unique solution of the constrained dual problem (3.66),*

$$\mathscr{A} = \{i : [\boldsymbol{\lambda}]_i > 0\} \cup \mathscr{E},$$

and $\mathsf{B}_{\mathscr{A}*}\mathsf{R}$ *is a full column rank matrix, then there is a unique triple* $(\widehat{\mathbf{x}}, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\alpha}})$ *such that* $(\widehat{\mathbf{x}}, \widehat{\boldsymbol{\lambda}})$ *solves the primal problem (3.44) and* $(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\alpha}})$ *solves the constrained dual problem (3.66). If* $\widehat{\boldsymbol{\lambda}}$ *is known, then*

$$\widehat{\boldsymbol{\alpha}} = (\mathsf{R}^T\mathsf{B}_{\mathscr{A}*}^T\mathsf{B}_{\mathscr{A}*}\mathsf{R})^{-1}\mathsf{R}^T\mathsf{B}_{\mathscr{A}*}^T\left(\mathsf{B}_{\mathscr{A}*}\mathsf{A}^+\mathsf{B}^T\widehat{\boldsymbol{\lambda}} - (\mathsf{B}_{\mathscr{A}*}\mathsf{A}^+\mathbf{b} - \mathbf{c}_{\mathscr{A}})\right) \tag{3.67}$$

and

$$\widehat{\mathbf{x}} = \mathsf{A}^+(\mathbf{b} - \mathsf{B}^T\widehat{\boldsymbol{\lambda}}) + \mathsf{R}\widehat{\boldsymbol{\alpha}}. \tag{3.68}$$

(iii) If B^T *and* $\mathsf{B}_{\mathscr{E}*}\mathsf{R}$ *are full column rank matrices, then there is a unique triple* $(\widehat{\mathbf{x}}, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\alpha}})$ *such that* $(\widehat{\mathbf{x}}, \widehat{\boldsymbol{\lambda}})$ *solves the primal problem (3.44) and* $(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\alpha}})$ *solves the constrained dual problem (3.66).*

Proof (i) Let B^T and BR be full column rank matrices. To show that there is a unique solution of (3.66), we examine the Hessian $\mathsf{BA}^+\mathsf{B}^T$ of θ. Let $\mathsf{R}^T\mathsf{B}^T\boldsymbol{\lambda} = \mathbf{o}$ and $\mathsf{BA}^+\mathsf{B}^T\boldsymbol{\lambda} = \mathbf{o}$. Using the definition of R, it follows that $\mathsf{B}^T\boldsymbol{\lambda} \in \mathrm{Im}\mathsf{A}$. Hence there is $\boldsymbol{\mu} \in \mathbb{R}^n$ such that

$$\mathsf{B}^T\boldsymbol{\lambda} = \mathsf{A}\boldsymbol{\mu}$$

and

$$\boldsymbol{\mu}^T\mathsf{A}\boldsymbol{\mu} = \boldsymbol{\mu}^T\mathsf{A}\mathsf{A}^+\mathsf{A}\boldsymbol{\mu} = \boldsymbol{\lambda}^T\mathsf{B}\mathsf{A}^+\mathsf{B}^T\boldsymbol{\lambda} = 0.$$

Thus $\boldsymbol{\mu} \in \mathrm{Ker}\mathsf{A}$ and

$$\mathsf{B}^T\boldsymbol{\lambda} = \mathsf{A}\boldsymbol{\mu} = \mathbf{o}.$$

Since we assume that B^T has independent columns, we conclude that $\boldsymbol{\lambda} = \mathbf{o}$. We have thus proved that the restriction of $\mathsf{BA}^+\mathsf{B}^T$ to $\mathrm{Ker}(\mathsf{R}^T\mathsf{B}^T)$ is positive definite, so that $\theta|\mathrm{Ker}\mathsf{R}^T\mathsf{B}^T$ is by Proposition 3.4 strictly convex, and it is easy to check that it is strictly convex on

$$\mathscr{U} = \{\boldsymbol{\lambda} \in \mathbb{R}^m : \ \mathsf{R}^T\mathsf{B}^T\boldsymbol{\lambda} = \mathsf{R}^T\mathbf{b}\}.$$

Since $\Omega_{BE} \neq \emptyset$ and $\Omega_{BE} \subseteq \mathscr{U}$, we have that θ is strictly convex on Ω_{BE}, and it follows by Proposition 3.3 that there is a unique solution $\widehat{\boldsymbol{\lambda}}$ of (3.66).
(ii) Let $\widehat{\boldsymbol{\lambda}}$ be a unique solution of problem (3.66). Since the solution satisfies the related KKT conditions, it follows that there is $\widehat{\boldsymbol{\alpha}}$ such that

$$\mathsf{B}_{\mathscr{A}*}\mathsf{A}^+\mathsf{B}^T\widehat{\boldsymbol{\lambda}} - (\mathsf{B}_{\mathscr{A}*}\mathsf{A}^+\mathbf{b} - \mathbf{c}_{\mathscr{A}}) - \mathsf{B}_{\mathscr{A}*}\mathsf{R}\widehat{\boldsymbol{\alpha}} = \mathbf{o}.$$

After multiplying on the left by $\mathsf{R}^T\mathsf{B}^T_{\mathscr{A}*}$ and simple manipulations, we get (3.67). The inverse exists and the solution $\widehat{\boldsymbol{\alpha}}$ is unique due to the uniqueness of $\widehat{\boldsymbol{\lambda}}$ and the assumption on the full column rank of $\mathsf{B}_{\mathscr{A}*}\mathsf{R}$.
(iii) If B^T and $\mathsf{B}_{\mathscr{E}*}\mathsf{R}$ are full column rank matrices, then BR is also a full column rank matrix. Hence, there is a unique solution $\widehat{\boldsymbol{\lambda}}$ of problem (3.66) by (i). Since $\mathscr{E} \subseteq \mathscr{A}$ and $\mathsf{B}_{\mathscr{E}*}\mathsf{R}$ has independent columns, it follows that $\mathsf{B}_{\mathscr{A}*}\mathsf{R}$ has also independent columns. Thus we can use (ii) to finish the proof. □

The reconstruction formula (3.67) can be modified in order to work whenever the dual problem has a solution $\overline{\boldsymbol{\lambda}}$. The resulting formula obtained by the analysis of the related KKT conditions then reads

$$\overline{\boldsymbol{\alpha}} = (\mathsf{R}^T\mathsf{B}^T_{\mathscr{A}*}\mathsf{B}_{\mathscr{A}*}\mathsf{R})^+\mathsf{R}^T\mathsf{B}^T_{\mathscr{A}*}\left(\mathsf{B}_{\mathscr{A}*}\mathsf{A}^+\mathsf{B}^T\overline{\boldsymbol{\lambda}} - (\mathsf{B}_{\mathscr{A}*}\mathsf{A}^+\mathbf{b} - \mathbf{c}_{\mathscr{A}})\right). \tag{3.69}$$

The duality theory can be illustrated on a problem to find the displacement $\mathbf{x}$ of an elastic body under traction $\mathbf{b}$. After the finite element discretization, we get a convex QP problem. We assume that the body is fixed on a part of the boundary in normal direction, so that the vector of nodal displacements satisfies $\mathsf{B}_{\mathscr{E}*}\mathbf{x} = \mathbf{c}_{\mathscr{E}}$ as in Fig. 3.9. Moreover, the body may not be allowed to penetrate an obstacle, so that $\mathsf{B}_{\mathscr{I}*}\mathbf{x} \leq \mathbf{c}_{\mathscr{I}}$ as in Fig. 3.8.

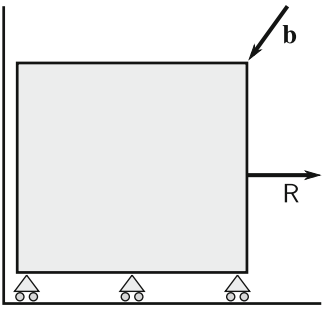

Fig. 3.8 Unique displacement

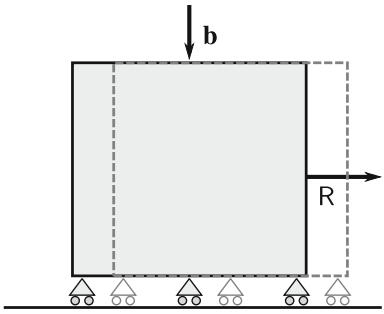

Fig. 3.9 Nonunique displacement

The displacement $\overline{\mathbf{x}}$ of the body in equilibrium is a minimizer of the convex energy function f. The Hessian A of f is positive semidefinite if the constraints admit rigid body motions. The Lagrange multipliers solve the dual problem. The condition $\mathsf{R}^T\mathbf{b} = \mathsf{R}^T\mathsf{B}^T\widehat{\boldsymbol{\lambda}}$ requires that the resulting forces are balanced in the directions of the rigid body motions and $\widehat{\boldsymbol{\lambda}}_{\mathscr{I}} \geq \mathbf{o}$ guarantees that the body is not glued to the obstacle. If the reaction forces $\mathsf{B}^T\widehat{\boldsymbol{\lambda}}$ determine the components of $\widehat{\boldsymbol{\lambda}}$, then $\widehat{\boldsymbol{\lambda}}$ is uniquely determined by the conditions of equilibrium. Notice that $\mathsf{B}^T\widehat{\boldsymbol{\lambda}}$ is always uniquely determined by the conditions of equilibrium. If no rigid body motion is possible due to the active constraints $\mathsf{B}_{\mathscr{A}*}\mathbf{x} = \mathbf{c}_{\mathscr{A}}$ as in Fig. 3.8, then the displacement $\mathbf{x}$ is uniquely determined. If this is not the case, then the displacement is determined up to some rigid body motion as in Fig. 3.9.

References

1. Bertsekas, D.P.: Nonlinear Optimization. Athena Scientific, Belmont (1999)
2. Nocedal, J., Wright, S.F.: Numerical Optimization. Springer, New York (2000)
3. Conn, A.R., Gould, N.I.M., Toint, Ph.L: Trust Region Methods. SIAM, Philadelphia (2000)
4. Bazaraa, M.S., Shetty, C.M., Sherali, H.D.: Nonlinear Programming, Theory and Algorithms, 2nd edn. Wiley, New York (1993)
5. Griva, I., Nash, S.G., Sofer, A.: Linear and Nonlinear Optimization. SIAM, Philadelphia (2009)
6. Frank, M., Wolfe, P.: An algorithm for quadratic programming. Naval Research Logistic Quarterly **3**, 95–110 (1956)
7. Dostál, Z.: On solvability of convex non-coercive quadratic programming problems. JOTA **143**(2), 413–416 (2009)

Chapter 4
Analysis

In this chapter, we first give a brief presentation of basic Sobolev spaces on Lipschitz domains and their boundaries; for more general results we refer to the books by Adams and Fournier [1], or McLean [2]. Then we review some ideas concerning semi-elliptic variational inequalities that provide an abstract framework for the formulation of contact problems. More information on variational inequalities can be found, e.g., in Lions and Stampacchia [3], Glowinski [4], and Glowinski, Lions, and Trémolier [5].

4.1 Sobolev Spaces

Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, denote a nonempty bounded Lipschitz domain with a boundary Γ. The space of all real functions that are Lebesgue measurable and quadratically integrable in Ω is denoted as $L^2(\Omega)$. We do not distinguish functions which differ on a set of zero measure. In $L^2(\Omega)$ we define the scalar product

$$(u, v) = (u, v)_{L^2(\Omega)} = \int_\Omega u\, v \, \mathrm{d}\Omega \tag{4.1}$$

and the norm

$$\|u\| = \|u\|_{L^2(\Omega)} = (u, u)^{1/2}.$$

The space $L^2(\Omega)$ with scalar product (4.1) is a Hilbert space. Note that there holds Hölder's inequality

$$\int_\Omega |u\, v| \, \mathrm{d}\Omega \le \|u\| \|v\| \quad \text{for all } u, v \in L^2(\Omega).$$

By $C^\infty(\overline{\Omega})$ we denote the space of all real functions with continuous derivatives of all orders with respect to all variables in Ω that are continuously extendable to its closure $\overline{\Omega}$.

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3_4

Let us define the Sobolev space

$$H^1(\Omega) = \left\{ u \in L^2(\Omega) : \frac{\partial u}{\partial x_i} \in L^2(\Omega) \text{ for } i = 1, \ldots, d \right\},$$

where the derivatives are considered in the weak sense. Note that $H^1(\Omega)$ can be equivalently defined as the completion of

$$\left(C^\infty(\overline{\Omega}), \| \cdot \|_{H^1(\Omega)} \right),$$

where

$$\|u\|_{H^1(\Omega)} = \left(\|u\|^2_{L^2(\Omega)} + |u|^2_{H^1(\Omega)} \right)^{1/2}$$

with

$$|u|^2_{H^1(\Omega)} = \|\nabla u\|^2.$$

It holds that $H^1(\Omega)$ is a Hilbert space with respect to the scalar product

$$(u, v)_{H^1(\Omega)} = (u, v)_{L^2(\Omega)} + (\nabla u, \nabla v)_{L^2(\Omega)}. \tag{4.2}$$

For a subset Γ_U of Γ with meas $\Gamma_U > 0$, we define the Sobolev space $H^1_0(\Omega, \Gamma_U)$ as the completion of

$$\left(C^\infty_0(\overline{\Omega}, \Gamma_U), \| \cdot \|_{H^1(\Omega)} \right),$$

where $C^\infty_0(\overline{\Omega}, \Gamma_U)$ contains all functions from $C^\infty(\overline{\Omega})$ vanishing on Γ_U. Note that $H^1_0(\Omega, \Gamma_U)$ is a Hilbert space with respect to the scalar product (4.2). In addition, due to the Friedrichs theorem, the functional $| \cdot |_{H^1(\Omega)}$ represents on $H^1_0(\Omega, \Gamma_U)$ an equivalent norm to $\| \cdot \|_{H^1(\Omega)}$. The following theorem defines the trace operator.

Theorem 4.1 (Trace Theorem) *There is a unique linear continuous mapping*

$$\gamma_0 : H^1(\Omega) \mapsto L^2(\Gamma) \tag{4.3}$$

satisfying

$$\gamma_0 u = u|\,\Gamma \quad \textit{for all } u \in C^\infty(\overline{\Omega}).$$

The function $\gamma_0 u \in L^2(\Gamma)$ is called the trace *of $u \in H^1(\Omega)$.*

Proof See, e.g., McLean [2]. □

Note that it can be shown that

$$H_0^1(\Omega, \Gamma_U) = \{u \in H^1(\Omega) : \gamma_0 u = 0 \text{ on } \Gamma_U\}.$$

4.2 Trace Spaces

Let us denote the trace space of $H^1(\Omega)$ by $H^{1/2}(\Gamma)$, i.e.,

$$H^{1/2}(\Gamma) = \gamma_0(H^1(\Omega)).$$

In $H^{1/2}(\Gamma)$ we introduce the *Sobolev–Slobodeckij scalar product*

$$(u, v)_{H^{1/2}(\Gamma)} = (u, v)_{L^2(\Gamma)} + \int_\Gamma \int_\Gamma \frac{(u(x) - u(y))(v(x) - v(y))}{\|x - y\|^d} \, \mathrm{d}\Gamma_x \, \mathrm{d}\Gamma_y \quad (4.4)$$

and the corresponding norm

$$\|v\|_{H^{1/2}(\Gamma)} = \left(\|v\|^2_{L^2(\Gamma)} + |v|^2_{H^{1/2}(\Gamma)}\right)^{1/2},$$

where

$$|v|^2_{H^{1/2}(\Gamma)} = \int_\Gamma \int_\Gamma \frac{(v(x) - v(y))^2}{\|x - y\|^d} \, \mathrm{d}\Gamma_x \, \mathrm{d}\Gamma_y.$$

Recall that $H^{1/2}(\Gamma)$ is a Hilbert space with respect to the scalar product (4.4), that there is $c_1 > 0$ such that

$$\|\gamma_0 u\|_{H^{1/2}(\Gamma)} \le c_1 \|u\|_{H^1(\Omega)} \quad \text{for all } u \in H^1(\Omega),$$

and that there exists $c_2 > 0$ and an extension operator $\varepsilon : H^{1/2}(\Gamma) \to H^1(\Omega)$ such that $\gamma_0 \varepsilon u = u$ and

$$\|\varepsilon u\|_{H^1(\Omega)} \le c_2 \|u\|_{H^{1/2}(\Gamma)} \quad \text{for all } u \in H^{1/2}(\Gamma).$$

The dual space to $H^{1/2}(\Gamma)$ with respect to the $L^2(\Gamma)$ scalar product is denoted by $H^{-1/2}(\Gamma)$ and the norm in $H^{-1/2}(\Gamma)$ is given by

$$\|w\|_{H^{-1/2}(\Gamma)} = \sup_{0 \neq v \in H^{1/2}(\Gamma)} \frac{|\langle w, v\rangle|}{\|v\|_{H^{1/2}(\Gamma)}},$$

where $\langle w, v\rangle = w(v)$ denotes the *duality pairing*. Notice that w is a bounded functional on L^2_Γ, so that by the Riesz theorem there is $\overline{w} \in L^2(\Gamma)$ such that

$$w(u) = (\overline{w}, u)_{L^2(\Gamma)}.$$

A key tool in the development of the boundary element method is the following Green formula. We shall formulate it for the functions from the space

$$H^1_\Delta(\Omega) = \{v \in H^1(\Omega) : \Delta v \in L^2(\Omega)\}.$$

Theorem 4.2 (Green's Theorem) *Let $u \in H^1_\Delta(\Omega)$ and $v \in H^1(\Omega)$.*

Then

$$(\nabla u, \nabla v)_{L^2(\Omega)} + (\Delta u, v)_{L^2(\Omega)} = \langle \gamma_1 u, \gamma_0 v \rangle , \tag{4.5}$$

where $\gamma_1 : H^1_\Delta(\Omega) \to H^{-1/2}(\Gamma)$ is the interior conormal derivative *defined by*

$$\gamma_1 u(\mathbf{x}) = \lim_{\Omega \ni \mathbf{y} \to \mathbf{x}} \frac{\partial}{\partial \mathbf{n}_\mathbf{x}} u(\mathbf{y}), \quad \mathbf{x} \in \Gamma. \tag{4.6}$$

Proof See, e.g., McLean [2]. □

4.3 Variational Inequalities

Let us review some basic ideas concerning semi-elliptic variational inequalities. Let V denote a real Hilbert space with a scalar product $(\cdot, \cdot)_V$ and the induced norm $\|\cdot\|_V$, let $|\cdot|_V$ be a seminorm on V, let $\mathcal{K} \subset V$ be a closed, convex, and nonempty set, let f be a bounded linear functional on V, i.e., $f \in V'$, and let a be a bilinear form on V.

Definition 4.1 A bilinear form $a : V \times V \to \mathbb{R}$ is said to be

- *bounded* on $\mathcal{K}$ if there is an $M > 0$ such that

$$|a(u, v)| \le M \|u\|_V \|v\|_V \quad \text{for all } u, v \in \mathcal{K},$$

- *symmetric* on $\mathcal{K}$ if

$$a(u, v) = a(v, u) \quad \text{for all } u, v \in \mathcal{K},$$

- *elliptic* on $\mathcal{K}$ if there is an $\alpha > 0$ such that

$$a(u, u) \ge \alpha \|u\|_V^2 \quad \text{for all } u \in \mathcal{K},$$

- *semi-elliptic* on $\mathcal{K}$ if there is an $\alpha > 0$ such that

$$a(u, u) \ge \alpha |u|_V^2 \quad \text{for all } u \in \mathcal{K}.$$

Definition 4.2 A functional $f : V \to \mathbb{R}$ is said to be

- *coercive* on $\mathscr{K}$ if
$$\left.\begin{array}{r} v \in \mathscr{K} \\ \|v\|_V \to \infty \end{array}\right\} \Rightarrow \; f(v) \to \infty,$$
- *convex* on $\mathscr{K}$ if for all $u, v \in \mathscr{K}$ and all $t \in (0, 1)$
$$f(tu + (1-t)v) \leq t\, f(u) + (1-t)\, f(v).$$

Theorem 4.3 *If a functional $f : V \to \mathbb{R}$ is continuous, coercive, and convex on $\mathscr{K}$, then there exists a solution of the minimization problem to find $u \in \mathscr{K}$ such that*

$$f(u) = \min\{f(v) : \; v \in \mathscr{K}\}.$$

Proof Let $\mathscr{B}_R$ be a closed, origin-centered ball of radius $R > 0$, i.e.,

$$\mathscr{B}_R = \{v \in V : \; \|v\|_V \leq R\}.$$

Then the coercivity of f on $\mathscr{K}$ yields the existence of a large enough $R > 0$ such that

$$q = \inf\{f(v) : \; v \in \mathscr{K}\} = \inf\{f(v) : \; v \in \mathscr{K} \cap \mathscr{B}_R\}.$$

Let us consider a sequence $\{u_n\} \subset \mathscr{K} \cap \mathscr{B}_R$ such that

$$f(u_n) \to q.$$

Since $\{u_n\}$ is bounded, there is a subsequence $\{u_{n_k}\}$ of $\{u_n\}$ and $u \in V$ satisfying

$$u_{n_k} \rightharpoonup u.$$

Let us recall that every closed convex set is weakly closed, i.e., $u \in \mathscr{K}$, and since every continuous convex functional is weakly lower semi-continuous on a closed convex set, we get

$$f(u) \leq \liminf f(u_{n_k}) = \lim f(u_{n_k}) = q \leq f(u).$$

□

We are concerned with the following *variational inequality*: find $u \in \mathscr{K}$ such that

$$a(u, v-u) \geq f(v-u) \quad \text{for all } v \in \mathscr{K}. \tag{4.7}$$

Note that if $\mathscr{K}$ is a subspace of V, then (4.7) is equivalent to the *variational equation* to find $u \in \mathscr{K}$ such that

$$a(u, v) = f(v) \quad \text{for all } v \in \mathscr{K}.$$

Theorem 4.4 (Generalization of the Lax–Milgram Theorem) *If a is bounded and elliptic on V, then there exists a unique solution $u \in \mathscr{K}$ of problem (4.7).*

Proof The proof can be found, e.g., in Glowinski [4], Stampacchia [6], and Glowinski, Lions, and Tremoliere [5]. Let us prove the uniqueness of the solution. Assume that both u_1 and u_2 solve the variational inequality (4.7), so that for all $v \in \mathscr{K}$ we have

$$a(u_1, v - u_1) \geq f(v - u_1) \quad \text{and} \quad a(u_2, v - u_2) \geq f(v - u_2).$$

Substituting u_2 and u_1 for v in the first and second inequality, respectively, and summing up the inequalities, we get

$$a(u_1 - u_2, u_1 - u_2) \leq 0.$$

Since a is elliptic on V, there is an $\alpha > 0$ such that

$$a(u_1 - u_2, u_1 - u_2) \geq \alpha \|u_1 - u_2\|_V^2.$$

Thus we get $u_1 = u_2$. □

Let us define the *energy functional* $q : V \to \mathbb{R}$ by

$$q(v) = \frac{1}{2} a(v, v) - f(v). \tag{4.8}$$

Proposition 4.1 *If a is symmetric and semi-elliptic on V, then the energy functional q defined by (4.8) is convex on V.*

Proof Let us show that for all $u, v \in V$ and all $t \in (0, 1)$

$$q(tu + (1 - t)v) \leq t\, q(u) + (1 - t) q(v).$$

Since f is linear on V, it is enough to prove that $v \mapsto a(v, v)$ is convex on V. Let $u, v \in V$ and $t \in (0, 1)$ be arbitrary. From the semi-ellipticity and symmetry of a on V we obtain

$$0 \leq a(u - v, u - v) = a(u, u) - 2\, a(u, v) + a(v, v). \tag{4.9}$$

Then, by (4.9), we get

$$\begin{aligned} a(tu + (1-t)v,\ tu + (1-t)v) &= \\ &= t^2 a(u,\ u) + 2t(1-t)\, a(u,\ v) + (1-t)^2 a(v,\ v) \\ &\le t^2 a(u,\ u) + t(1-t)\,[a(u,\ u) + a(v,\ v)] + (1-t)^2 a(v,\ v) \\ &= t\, a(u,\ u) + (1-t)\, a(v,\ v), \end{aligned}$$

which completes the proof. □

Theorem 4.5 *If a is bounded, symmetric, and semi-elliptic on V, then problem (4.7) is equivalent to the minimization problem to find $u \in \mathscr{K}$ such that*

$$q(u) = \min\{q(v) :\ v \in \mathscr{K}\}. \tag{4.10}$$

Proof Suppose $u \in \mathscr{K}$ is a solution of (4.7). Let us pick any $v \in \mathscr{K}$ and put $z = v - u \in V$. Then

$$a(u, z) \ge f(z)$$

and due to the symmetry and semi-ellipticity of a on V

$$\begin{aligned} q(v) = q(z+u) &= \frac{1}{2}a(z+u, z+u) - f(z+u) \\ &= \frac{1}{2}a(z,z) + a(u,z) + \frac{1}{2}a(u,u) - f(z) - f(u) \\ &= q(u) + \frac{1}{2}a(z,z) + (a(u,z) - f(z)) \\ &\ge q(u). \end{aligned}$$

Conversely, assume that $u \in \mathscr{K}$ minimizes the energy functional q on $\mathscr{K}$. If we pick an arbitrary $v \in \mathscr{K}$, then

$$\phi(t) = q((1-t)u + tv) \ge q(u) = \phi(0) \quad \text{for all } t \in [0,1].$$

Now let us take a closer look at the function ϕ. The symmetry of a on V yields

$$\phi(t) = \frac{1}{2}(1-t)^2 a(u,u) + (1-t)\, t a(u,v) + \frac{1}{2}t^2 a(v,v) - (1-t) f(u) - t f(v)$$

for all $t \in [0,1]$, and therefore

$$\phi'_+(0) = -a(u,u) + a(u,v) + f(u) - f(v) = a(u, v-u) - f(v-u).$$

Since $\phi'_+(0) \ge 0$, the proof is finished. □

Remark 4.1 Let us sketch the proof of Theorem 4.4 using an additional assumption of the symmetry of a on V. Due to Theorem 4.5, to prove the solvability, it is enough to show that the energy functional q defined by (4.8) gets its minimum on $\mathscr{K}$. By Proposition 4.1 we know that q is convex on V. Continuity of both a and f on V implies continuity of q on V and, moreover, it can be easily shown that particularly due to elipticity of a on V the energy functional q is coercive on V. Thus, by Theorem 4.3, we get that problem (4.10) is solvable.

Remark 4.2 If the assumptions of Proposition 4.1 are satisfied, q is convex on V. If, moreover, a is continuous on V, so is q. Thus if we want to use Theorem 4.3 in order to obtain the solvability of minimization problem (4.10) (and the solvability of variational inequality (4.7)), it is enough to prove coercivity of q on $\mathscr{K}$.

References

1. Adams, R.A., Fournier, J.J.F.: Sobolev Spaces, 2nd edn. Academic Press, New York (2003)
2. McLean, W.: Strongly Elliptic Systems and Boundary Integral Equations. Cambridge University Press, Cambridge (2000)
3. Lions, J.-L., Stampacchia, G.: Variational inequalities. Comm. Pure Appl. Math. **20**, 493–519 (1967)
4. Glowinski, R.: Numerical Methods for Nonlinear Variational Problems. Springer Series in Computational Physics. Springer, Berlin (1984)
5. Glowinski, R., Lions, J.L., Tremolieres, R.: Numerical Analysis of Variational Inequalities. North-Holland, Amsterdam (1981)
6. Stampacchia, G.: Formes bilineaires coercitives sur les ensembles convexes. C. r. hebd. séances Acad. sci. **258**, 4413–4416 (1964)

Part II
Optimal QP and QCQP Algorithms

Chapter 5
Conjugate Gradients

We begin our development of scalable algorithms for contact problems by the description of the *conjugate gradient method* for solving

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}), \tag{5.1}$$

where $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathsf{A}\mathbf{x} - \mathbf{x}^T\mathbf{b}$, $\mathbf{b}$ is a given column n-vector, and A is an $n \times n$ SPD matrix. We are interested especially in the problems with n large and A sparse and reasonably conditioned.

As we have already seen in Sect. 3.2.2, problem (5.1) is equivalent to the solution of a system of linear equations $\mathsf{A}\mathbf{x} = \mathbf{b}$, but our main goal here is the solution of auxiliary minimization problems generated by the algorithms for solving constrained QP problems arising from the discretization of contact problems. Here, we view the conjugate gradient (CG) method as an *iterative method*, which generates improving approximations to the solution of (5.1) at each step. The cost of one step of the CG method is dominated by the cost of the multiplication of a vector by the matrix A, which is proportional to the number of nonzero entries of A or its sparse representation. The memory requirements are typically dominated by the cost of the storage of A. In our applications, the matrices come in the form of a product of sparse matrices so that the cost of matrix–vector multiplications and storage requirements increases proportionally to the number of unknown variables.

The rate of convergence of the CG method depends on the distribution of the spectrum of A and can be improved by a problem- dependent *preconditioning*. For the development of scalable algorithms for contact problems, it is important that there are well-established standard preconditioners for the solution of the problems arising from the application of domain decomposition methods to the problems of elasticity that can reduce the conditioning of the Hessian matrix of the discretized elastic energy function.

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3_5

5.1 First Observations

The conjugate gradient method is based on simple observations. Let us start with examining the first one, namely, that it is possible to reduce the solution of (5.1) to the solution of a sequence of one-dimensional problems.

Let $\mathsf{A} \in \mathbb{R}^{n\times n}$ be an SPD matrix and let us assume that there are nonzero n-vectors $\mathbf{p}^1, \ldots, \mathbf{p}^n$ such that

$$(\mathbf{p}^i, \mathbf{p}^j)_{\mathsf{A}} = (\mathbf{p}^i)^T \mathsf{A}\mathbf{p}^j = 0 \ \text{ for } \ i \neq j.$$

We call such vectors A-*conjugate* or briefly *conjugate*. Specializing the arguments of Sect. 2.6, we get that $\mathbf{p}^1, \ldots, \mathbf{p}^n$ are independent. Thus $\mathbf{p}^1, \ldots, \mathbf{p}^n$ form the basis of $\mathbb{R}^n$ and any $\mathbf{x} \in \mathbb{R}^n$ can be written in the form

$$\mathbf{x} = \xi_1 \mathbf{p}^1 + \cdots + \xi_n \mathbf{p}^n.$$

Substituting into f and using the conjugacy results in

$$\begin{aligned} f(\mathbf{x}) &= \left(\frac{1}{2}\xi_1^2 (\mathbf{p}^1)^T \mathsf{A}\mathbf{p}^1 - \xi_1 \mathbf{b}^T \mathbf{p}^1\right) + \cdots + \left(\frac{1}{2}\xi_n^2 (\mathbf{p}^n)^T \mathsf{A}\mathbf{p}^n - \xi_n \mathbf{b}^T \mathbf{p}^n\right) \\ &= f(\xi_1 \mathbf{p}^1) + \cdots + f(\xi_n \mathbf{p}^n). \end{aligned}$$

Thus

$$f(\widehat{\mathbf{x}}) = \min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = \min_{\xi_1\in\mathbb{R}} f(\xi_1 \mathbf{p}^1) + \cdots + \min_{\xi_n\in\mathbb{R}} f(\xi_n \mathbf{p}^n).$$

We have thus managed to decompose the original problem (5.1) into n one-dimensional problems. Since

$$\left.\frac{\mathrm{d} f\left(\xi \mathbf{p}^i\right)}{\mathrm{d}\xi}\right|_{\xi_i} = \xi_i (\mathbf{p}^i)^T \mathsf{A}\mathbf{p}^i - \mathbf{b}^T \mathbf{p}^i = 0,$$

the solution $\widehat{\mathbf{x}}$ of (5.1) is given by

$$\widehat{\mathbf{x}} = \xi_1 \mathbf{p}^1 + \cdots + \xi_n \mathbf{p}^n, \quad \xi_i = \mathbf{b}^T \mathbf{p}^i / (\mathbf{p}^i)^T \mathsf{A}\mathbf{p}^i, \ i = 1, \ldots, n. \tag{5.2}$$

The second observation concerns effective generation of a conjugate basis. Let us recall how to generate conjugate directions with the *Gram–Schmidt procedure*. Assuming that $\mathbf{p}^1, \ldots, \mathbf{p}^k$ are nonzero conjugate directions, $1 \leq k < n$, let us examine how to use $\mathbf{h}^k \notin \mathrm{Span}\{\mathbf{p}^1, \ldots, \mathbf{p}^k\}$ to generate a new member $\mathbf{p}^{k+1}$ in the form

$$\mathbf{p}^{k+1} = \mathbf{h}^k + \beta_{k1} \mathbf{p}^1 + \cdots + \beta_{kk} \mathbf{p}^k. \tag{5.3}$$

Since $\mathbf{p}^{k+1}$ should be conjugate to $\mathbf{p}^1, \ldots, \mathbf{p}^k$, we get

$$\begin{aligned} 0 &= (\mathbf{p}^i)^T \mathsf{A}\mathbf{p}^{k+1} = (\mathbf{p}^i)^T \mathsf{A}\mathbf{h}^k + \beta_{k1}(\mathbf{p}^i)^T \mathsf{A}\mathbf{p}^1 + \cdots + \beta_{kk}(\mathbf{p}^i)^T \mathsf{A}\mathbf{p}^k \\ &= (\mathbf{p}^i)^T \mathsf{A}\mathbf{h}^k + \beta_{ki}(\mathbf{p}^i)^T \mathsf{A}\mathbf{p}^i, \quad i = 1, \ldots, k. \end{aligned}$$

Thus

$$\beta_{ki} = -\frac{(\mathbf{p}^i)^T \mathsf{A}\mathbf{h}^k}{(\mathbf{p}^i)^T \mathsf{A}\mathbf{p}^i}, \ i = 1, \ldots, k. \tag{5.4}$$

Obviously

$$\mathrm{Span}\{\mathbf{p}^1, \ldots, \mathbf{p}^{k+1}\} = \mathrm{Span}\{\mathbf{p}^1, \ldots, \mathbf{p}^k, \mathbf{h}^k\}.$$

Therefore, given any independent vectors $\mathbf{h}^0, \ldots, \mathbf{h}^{k-1}$, we can start from $\mathbf{p}^1 = \mathbf{h}^0$ and use (5.3) and (5.4) to construct a set of mutually A-conjugate directions $\mathbf{p}^1, \ldots, \mathbf{p}^k$ such that

$$\mathrm{Span}\{\mathbf{h}^0, \ldots, \mathbf{h}^{i-1}\} = \mathrm{Span}\{\mathbf{p}^1, \ldots, \mathbf{p}^i\}, \ i = 1, \ldots, k.$$

For $\mathbf{h}^0, \ldots, \mathbf{h}^{k-1}$ arbitrary, the construction is increasingly expensive as it requires both the storage for the vectors $\mathbf{p}^1, \ldots, \mathbf{p}^k$ and heavy calculations including evaluation of $k(k+1)/2$ scalar products. However, it turns out that we can adapt the procedure so that it generates very efficiently the conjugate basis of the *Krylov spaces*

$$\mathscr{K}^k = \mathscr{K}^k(\mathsf{A}, \mathbf{g}^0) = \mathrm{Span}\{\mathbf{g}^0, \mathsf{A}\mathbf{g}^0, \ldots, \mathsf{A}^{k-1}\mathbf{g}^0\}, \quad k = 1, \ldots, n,$$

with $\mathbf{g}^0 = \mathsf{A}\mathbf{x}^0 - \mathbf{b}$ defined by a suitable initial vector $\mathbf{x}^0$ and $\mathscr{K}^0 = \{\mathbf{o}\}$. The powerful method is again based on a few simple observations.

First assume that $\mathbf{p}^1, \ldots, \mathbf{p}^i$ form a conjugate basis of $\mathscr{K}^i$, $i = 1, \ldots, k$, and observe that if $\mathbf{x}^k$ denotes the minimizer of f on $\mathbf{x}^0 + \mathscr{K}^k$, then by Proposition 3.6(i) the gradient $\mathbf{g}^k = \nabla f(\mathbf{x}^k)$ is orthogonal to the Krylov space $\mathscr{K}^k$, that is,

$$(\mathbf{g}^k)^T \mathbf{x} = 0 \ \text{ for any } \ \mathbf{x} \in \mathscr{K}^k.$$

In particular, if $\mathbf{g}^k \neq \mathbf{o}$, then

$$\mathbf{g}^k \notin \mathscr{K}^k.$$

Since $\mathbf{g}^k \in \mathscr{K}^{k+1}$, we can use (5.3) with $\mathbf{h}^k = \mathbf{g}^k$ to expand any conjugate basis of $\mathscr{K}^k$ to the conjugate basis of $\mathscr{K}^{k+1}$. Obviously

$$\mathscr{K}^k(\mathsf{A}, \mathbf{g}^0) = \mathrm{Span}\{\mathbf{g}^0, \ldots, \mathbf{g}^{k-1}\}.$$

Next observe that for any $\mathbf{x} \in \mathscr{K}^{k-1}$ and $k \geq 1$

$$\mathsf{A}\mathbf{x} \in \mathscr{K}^k,$$

or briefly $\mathsf{A}\mathscr{K}^{k-1} \subseteq \mathscr{K}^k$. Since $\mathbf{p}^i \in \mathscr{K}^i \subseteq \mathscr{K}^{k-1},\ i = 1, \ldots, k-1$, we have

$$(\mathsf{A}\mathbf{p}^i)^T \mathbf{g}^k = (\mathbf{p}^i)^T \mathsf{A}\mathbf{g}^k = 0,\ i = 1, \ldots, k-1.$$

It follows that

$$\beta_{ki} = -\frac{(\mathbf{p}^i)^T \mathsf{A}\mathbf{g}^k}{(\mathbf{p}^i)^T \mathsf{A}\mathbf{p}^i} = 0,\ i = 1, \ldots, k-1.$$

Summing up, if we have a set of such conjugate vectors $\mathbf{p}^1, \ldots, \mathbf{p}^k$ that

$$\mathrm{Span}\{\mathbf{p}^1, \ldots, \mathbf{p}^i\} = \mathscr{K}^i,\ i = 1, \ldots k,$$

then the formula (5.3) applied to $\mathbf{p}^1, \ldots, \mathbf{p}^k$ and $\mathbf{h}^k = \mathbf{g}^k$ simplifies to

$$\mathbf{p}^{k+1} = \mathbf{g}^k + \beta_k \mathbf{p}^k \tag{5.5}$$

with

$$\beta_k = \beta_{kk} = -\frac{(\mathbf{p}^k)^T \mathsf{A}\mathbf{g}^k}{(\mathbf{p}^k)^T \mathsf{A}\mathbf{p}^k}. \tag{5.6}$$

Finally, observe that the orthogonality of $\mathbf{g}^k$ to $\mathrm{Span}\{\mathbf{p}^1, \ldots, \mathbf{p}^k\}$ and (5.5) imply that

$$\|\mathbf{p}^{k+1}\| \geq \|\mathbf{g}^k\|. \tag{5.7}$$

In particular, if $\mathbf{g}^{k-1} \neq \mathbf{o}$, then $\mathbf{p}^k \neq \mathbf{o}$, so the formula (5.6) is well defined provided $\mathbf{g}^{k-1} \neq \mathbf{o}$.

5.2 Conjugate Gradient Method

In the previous two sections, we have found that the conjugate directions can be used to reduce the minimization of any convex quadratic function to the solution of a sequence of one-dimensional problems, and that the conjugate directions can be generated very efficiently. The famous *conjugate gradient (CG) method* just puts these two observations together.

The algorithm starts from an initial guess $\mathbf{x}^0$, $\mathbf{g}^0 = \mathsf{A}\mathbf{x}^0 - \mathbf{b}$, and $\mathbf{p}^1 = \mathbf{g}^0$. If $\mathbf{x}^{k-1}$ and $\mathbf{g}^{k-1}$ are given, $k \geq 1$, it first checks if $\mathbf{x}^{k-1}$ is the solution. If not, then the algorithm generates

$$\mathbf{x}^k = \mathbf{x}^{k-1} - \alpha_k \mathbf{p}^k \quad \text{with} \quad \alpha_k = (\mathbf{g}^{k-1})^T \mathbf{p}^k / (\mathbf{p}^k)^T \mathsf{A}\mathbf{p}^k \tag{5.8}$$

and

$$\begin{aligned} \mathbf{g}^k &= \mathsf{A}\mathbf{x}^k - \mathbf{b} = \mathsf{A}\left(\mathbf{x}^{k-1} - \alpha_k \mathbf{p}^k\right) - \mathbf{b} = \left(\mathsf{A}\mathbf{x}^{k-1} - \mathbf{b}\right) - \alpha_k \mathsf{A}\mathbf{p}^k \\ &= \mathbf{g}^{k-1} - \alpha_k \mathsf{A}\mathbf{p}^k. \end{aligned} \tag{5.9}$$

Finally the new conjugate direction $\mathbf{p}^{k+1}$ is generated by (5.5) and (5.6).

The decision if $\mathbf{x}^{k-1}$ is an acceptable solution is typically based on the value of $\|\mathbf{g}^{k-1}\|$, so the norm of the gradient must be evaluated at each step. It turns out that the norm can also be used to replace the scalar products involving the gradient in the definition of α_k and β_k. To find the formulae, let us replace k in (5.5) by $k-1$ and multiply the resulting identity by $(\mathbf{g}^{k-1})^T$. Using the orthogonality, we get

$$(\mathbf{g}^{k-1})^T \mathbf{p}^k = \|\mathbf{g}^{k-1}\|^2 + \beta_{k-1} (\mathbf{g}^{k-1})^T \mathbf{p}^{k-1} = \|\mathbf{g}^{k-1}\|^2, \tag{5.10}$$

so by (5.8)

$$\alpha_k = \frac{\|\mathbf{g}^{k-1}\|^2}{(\mathbf{p}^k)^T \mathsf{A}\mathbf{p}^k}. \tag{5.11}$$

To find an alternative formula for β_k, notice that $\alpha_k > 0$ for $\mathbf{g}^{k-1} \neq \mathbf{o}$ and that by (5.9)

$$\mathsf{A}\mathbf{p}^k = \frac{1}{\alpha_k}(\mathbf{g}^{k-1} - \mathbf{g}^k),$$

so that

$$\alpha_k (\mathbf{g}^k)^T \mathsf{A}\mathbf{p}^k = (\mathbf{g}^k)^T (\mathbf{g}^{k-1} - \mathbf{g}^k) = -\|\mathbf{g}^k\|^2$$

and

$$\beta_k = -\frac{(\mathbf{p}^k)^T \mathsf{A}\mathbf{g}^k}{(\mathbf{p}^k)^T \mathsf{A}\mathbf{p}^k} = \frac{\|\mathbf{g}^k\|^2}{\alpha_k (\mathbf{p}^k)^T \mathsf{A}\mathbf{p}^k} = \frac{\|\mathbf{g}^k\|^2}{\|\mathbf{g}^{k-1}\|^2}. \tag{5.12}$$

The complete CG method is presented as Algorithm 5.1.

Algorithm 5.1 Conjugate gradient method (CG).

Given a symmetric positive definite matrix $\mathsf{A} \in \mathbb{R}^{n\times n}$ *and* $\mathbf{b} \in \mathbb{R}^n$.

Step 0. {*Initialization.*}

Choose $\mathbf{x}^0 \in \mathbb{R}^n$, *set* $\mathbf{g}^0 = \mathsf{A}\mathbf{x}^0 - \mathbf{b}$, $\mathbf{p}^1 = \mathbf{g}^0$, $k = 1$

Step 1. {*Conjugate gradient loop.* }

while $\|\mathbf{g}^{k-1}\| > 0$

$\alpha_k = \|\mathbf{g}^{k-1}\|^2/(\mathbf{p}^k)^T\mathsf{A}\mathbf{p}^k$

$\mathbf{x}^k = \mathbf{x}^{k-1} - \alpha_k\mathbf{p}^k$

$\mathbf{g}^k = \mathbf{g}^{k-1} - \alpha_k\mathsf{A}\mathbf{p}^k$

$\beta_k = \|\mathbf{g}^k\|^2/\|\mathbf{g}^{k-1}\|^2 = -(\mathsf{A}\mathbf{p}^k)^T\mathbf{g}^k/\left((\mathbf{p}^k)^T\mathsf{A}\mathbf{p}^k\right)$

$\mathbf{p}^{k+1} = \mathbf{g}^k + \beta_k\mathbf{p}^k$

$k = k + 1$

end while

Step 2. {*Return the solution.*}

$\widehat{\mathbf{x}} = \mathbf{x}^k$

Each step of the CG method can be implemented with just one matrix–vector multiplication. This multiplication by the Hessian matrix A typically dominates the cost of the step. Only one generation of vectors $\mathbf{x}^k$, $\mathbf{p}^k$, and $\mathbf{g}^k$ is typically stored, so the memory requirements are modest.

Let us recall that the algorithm finds at each step the minimizer $\mathbf{x}^k$ of f on $\mathbf{x}^0 + \mathscr{K}^k = \mathbf{x}^0 + \mathscr{K}^k(\mathsf{A}, \mathbf{g}^0)$ and expands the conjugate basis of $\mathscr{K}^k$ to that of $\mathscr{K}^{k+1}$ provided $\mathbf{g}^k \neq \mathbf{o}$. Since the dimension of $\mathscr{K}^k$ is less than or equal to k, it follows that for some $k \leq n$

$$\mathscr{K}^k = \mathscr{K}^{k+1}.$$

Since $\mathbf{g}^k \in \mathscr{K}^{k+1}$ and $\mathbf{g}^k$ is orthogonal to $\mathscr{K}^k$, Algorithm 5.1 implemented in the exact arithmetics finds the solution $\widehat{\mathbf{x}}$ of (5.1) in at most n steps. We can sum up the most important properties of Algorithm 5.1 into the following theorem.

Theorem 5.1 *Let* $\{\mathbf{x}^k\}$ *be generated by Algorithm 5.1 to find the solution* $\widehat{\mathbf{x}}$ *of (5.1) starting from* $\mathbf{x}^0 \in \mathbb{R}^n$. *Then the algorithm is well defined and there is* $k \leq n$ *such that* $\mathbf{x}^k = \widehat{\mathbf{x}}$. *Moreover, the following statements hold for* $i = 1, \ldots, k$*:*

(i) $f(\mathbf{x}^i) = \min\{f(\mathbf{x}) : \mathbf{x} \in \mathbf{x}^0 + \mathscr{K}^i(\mathsf{A}, \mathbf{g}^0)\}$.
(ii) $\|\mathbf{p}^{i+1}\| \geq \|\mathbf{g}^i\|$.
(iii) $(\mathbf{g}^i)^T\mathbf{g}^j = 0$ *for* $i \neq j$.
(iv) $(\mathbf{p}^i)^T\mathsf{A}\mathbf{p}^j = 0$ *for* $i \neq j$.
(v) $\mathscr{K}^i(\mathsf{A}, \mathbf{g}^0) = \mathrm{Span}\{\mathbf{g}^0, \ldots, \mathbf{g}^{i-1}\} = \mathrm{Span}\{\mathbf{p}^1, \ldots, \mathbf{p}^i\}$.

It is usually sufficient to find $\mathbf{x}^k$ such that $\|\mathbf{g}^k\|$ is small. For example, given a small $\varepsilon > 0$, we can consider $\mathbf{g}^k$ small if

$$\|\mathbf{g}^k\| \leq \varepsilon\|\mathbf{b}\|.$$

Then $\widetilde{\mathbf{x}} = \mathbf{x}^k$ is an approximate solution which satisfies

$$\|\mathsf{A}(\widetilde{\mathbf{x}} - \widehat{\mathbf{x}})\| \le \varepsilon \|\mathbf{b}\|, \quad \|\widetilde{\mathbf{x}} - \widehat{\mathbf{x}}\| \le \varepsilon \lambda_{\min}(\mathsf{A})^{-1},$$

where $\lambda_{\min}(\mathsf{A})$ denotes the least eigenvalue of A. It is easy to check that the approximate solution $\widetilde{\mathbf{x}}$ solves the perturbed problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \widetilde{f}(\mathbf{x}), \quad \widetilde{f}(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathsf{A}\mathbf{x} - \widetilde{\mathbf{b}}^T\mathbf{x}, \quad \widetilde{\mathbf{b}} = \mathbf{b} + \mathbf{g}^k.$$

What is "small" depends on the problem solved. To keep our exposition general, we shall often not specify the test in what follows. Of course $\mathbf{g}^k = \mathbf{o}$ is always considered small.

5.3 Rate of Convergence

Although the CG method finds the exact solution $\widehat{\mathbf{x}}$ of (5.1) in a number of steps which does not exceed the dimension of the problem by Theorem 5.1, it turns out that it can often produce a sufficiently accurate approximation $\widetilde{\mathbf{x}}$ of $\widehat{\mathbf{x}}$ in a much smaller number of steps than required for exact termination. This observation suggests that the CG method may also be considered as an iterative method. In this section we present the results which substantiate this claim and help us to identify the favorable cases.

Let us denote the *solution error* as

$$\mathbf{e} = \mathbf{e}(\mathbf{x}) = \mathbf{x} - \widehat{\mathbf{x}}$$

and observe that

$$\mathbf{g}(\widehat{\mathbf{x}}) = \mathsf{A}\widehat{\mathbf{x}} - \mathbf{b} = \mathbf{o}.$$

It follows that

$$\mathbf{g}^k = \mathsf{A}\mathbf{x}^k - \mathbf{b} = \mathsf{A}\mathbf{x}^k - \mathsf{A}\widehat{\mathbf{x}} = \mathsf{A}(\mathbf{x}^k - \widehat{\mathbf{x}}) = \mathsf{A}\mathbf{e}^k,$$

so in particular

$$\mathscr{K}^k(\mathsf{A}, \mathbf{g}^0) = \mathrm{Span}\{\mathbf{g}^0, \mathsf{A}\mathbf{g}^0, \ldots, \mathsf{A}^{k-1}\mathbf{g}^0\} = \mathrm{Span}\{\mathsf{A}\mathbf{e}^0, \ldots, \mathsf{A}^k\mathbf{e}^0\}.$$

We start our analysis of the solution error by using the Taylor expansion (3.4) to obtain the identity

$$\begin{aligned} f(\mathbf{x}) - f(\widehat{\mathbf{x}}) &= f(\widehat{\mathbf{x}} + (\mathbf{x} - \widehat{\mathbf{x}})) - f(\widehat{\mathbf{x}}) \\ &= f(\widehat{\mathbf{x}}) + \mathbf{g}(\widehat{\mathbf{x}})^T(\mathbf{x} - \widehat{\mathbf{x}}) + \frac{1}{2}\|\mathbf{x} - \widehat{\mathbf{x}}\|_{\mathsf{A}}^2 - f(\widehat{\mathbf{x}}) \\ &= \frac{1}{2}\|\mathbf{x} - \widehat{\mathbf{x}}\|_{\mathsf{A}}^2 = \frac{1}{2}\|\mathbf{e}\|_{\mathsf{A}}^2. \end{aligned}$$

Combining the latter identity with Theorem 5.1, we get

$$\begin{aligned} \|\mathbf{e}^k\|_A^2 &= 2\left(f(\mathbf{x}^k) - f(\widehat{\mathbf{x}})\right) = \min_{\mathbf{x}\in\mathbf{x}^0+\mathscr{K}^k(\mathsf{A},\mathbf{g}^0)} 2\left(f(\mathbf{x}) - f(\widehat{\mathbf{x}})\right) \\ &= \min_{\mathbf{x}\in\mathbf{x}^0+\mathscr{K}^k(\mathsf{A},\mathbf{g}^0)} \|\mathbf{x} - \widehat{\mathbf{x}}\|_{\mathsf{A}}^2 = \min_{\mathbf{x}\in\mathbf{x}^0+\mathscr{K}^k(\mathsf{A},\mathbf{g}^0)} \|\mathbf{e}(\mathbf{x})\|_{\mathsf{A}}^2. \end{aligned}$$

Since any $\mathbf{x} \in \mathbf{x}^0 + \mathscr{K}^k(\mathsf{A}, \mathbf{g}^0)$ may be written in the form

$$\mathbf{x} = \mathbf{x}^0 + \xi_1\mathbf{g}^0 + \xi_2\mathsf{A}\mathbf{g}^0 + \cdots + \xi_k\mathsf{A}^{k-1}\mathbf{g}^0 = \mathbf{x}^0 + \xi_1\mathsf{A}\mathbf{e}^0 + \cdots + \xi_k\mathsf{A}^k\mathbf{e}^0,$$

it follows that

$$\mathbf{x} - \widehat{\mathbf{x}} = \mathbf{e}^0 + \xi_1\mathsf{A}\mathbf{e}^0 + \cdots + \xi_k\mathsf{A}^k\mathbf{e}^0 = p(\mathsf{A})\mathbf{e}^0,$$

where p denotes the polynomial defined for any $x \in \mathbb{R}$ by

$$p(x) = 1 + \xi_1 x + \xi_2 x^2 + \cdots + \xi_k x^k.$$

Thus denoting by $\mathscr{P}^k$ the set of all kth degree polynomials p which satisfy $p(0) = 1$, we have

$$\|\mathbf{e}^k\|_{\mathsf{A}}^2 = \min_{\mathbf{x}\in\mathbf{x}^0+\mathscr{K}^k(\mathsf{A},\mathbf{g}^0)} \|\mathbf{e}(\mathbf{x})\|_{\mathsf{A}}^2 = \min_{p\in\mathscr{P}^k} \|p(\mathsf{A})\mathbf{e}^0\|_{\mathsf{A}}^2. \tag{5.13}$$

We shall now derive a bound on the expression on the right-hand side of (5.13) that depends on the spectrum of A but is independent of the initial error $\mathbf{e}^0$. Let a spectral decomposition of A be written as $\mathsf{A} = \mathsf{U}\mathsf{D}\mathsf{U}^T$, where U is an orthogonal matrix and $\mathsf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix defined by the eigenvalues of A. Since A is assumed to be positive definite, the square root of A is well defined by

$$\mathsf{A}^{\frac{1}{2}} = \mathsf{U}\mathsf{D}^{\frac{1}{2}}\mathsf{U}^T.$$

Using $p(\mathsf{A}) = \mathsf{U}p(\mathsf{D})\mathsf{U}^T$, it is also easy to check that

$$\mathsf{A}^{\frac{1}{2}}p(\mathsf{A}) = p(\mathsf{A})\mathsf{A}^{\frac{1}{2}}.$$

Moreover, for any vector $\mathbf{v} \in \mathbb{R}^n$

$$\|\mathbf{v}\|_{\mathsf{A}}^2 = \mathbf{v}^T\mathsf{A}\mathbf{v} = \mathbf{v}^T\mathsf{A}^{\frac{1}{2}}\mathsf{A}^{\frac{1}{2}}\mathbf{v} = (\mathsf{A}^{\frac{1}{2}}\mathbf{v})^T\mathsf{A}^{\frac{1}{2}}\mathbf{v} = \|\mathsf{A}^{\frac{1}{2}}\mathbf{v}\|^2.$$

Using the latter identities (5.13), and the properties of norms, we get

$$\begin{aligned}\|\mathbf{e}^k\|_A^2 &= \min_{p\in\mathscr{P}^k} \|p(\mathsf{A})\mathbf{e}^0\|_{\mathsf{A}}^2 = \min_{p\in\mathscr{P}^k} \|\mathsf{A}^{\frac{1}{2}} p(\mathsf{A})\mathbf{e}^0\|^2 = \min_{p\in\mathscr{P}^k} \|p(\mathsf{A})\mathsf{A}^{\frac{1}{2}}\mathbf{e}^0\|^2 \\ &\le \min_{p\in\mathscr{P}^k} \|p(\mathsf{A})\|^2 \|\mathsf{A}^{\frac{1}{2}}\mathbf{e}^0\|^2 = \min_{p\in\mathscr{P}^k} \|p(\mathsf{D})\|^2 \|\mathbf{e}^0\|_{\mathsf{A}}^2.\end{aligned}$$

Since

$$\|p(\mathsf{D})\| = \max_{i\in\{1,\dots,n\}} |p(\lambda_i)|,$$

we can write

$$\|\mathbf{e}^k\|_{\mathsf{A}} \le \min_{p\in\mathscr{P}^k} \max_{i\in\{1,\dots,n\}} |p(\lambda_i)|\, \|\mathbf{e}^0\|_{\mathsf{A}}. \tag{5.14}$$

The estimate (5.14) reduces the analysis of convergence of the CG method to the analysis of approximation of zero function on $\sigma(\mathsf{A})$ of A by a kth degree polynomial with the value one at origin. For example, if $\sigma(\mathsf{A})$ is clustered around a single point ξ, then the minimization by the CG should be very effective because $|(1 - x/\xi)^k|$ is small near ξ. We shall use (5.14) to get a "global" estimate of the rate of convergence of the CG method in terms of the condition number of A.

Theorem 5.2 *Let $\{\mathbf{x}^k\}$ be generated by Algorithm 5.1 to find the solution $\widehat{\mathbf{x}}$ of (5.1) starting from $\mathbf{x}^0 \in \mathbb{R}^n$. Then the error*

$$\mathbf{e}^k = \mathbf{x}^k - \widehat{\mathbf{x}}$$

satisfies

$$\|\mathbf{e}^k\|_{\mathsf{A}} \le 2\left(\frac{\sqrt{\kappa(\mathsf{A})} - 1}{\sqrt{\kappa(\mathsf{A})} + 1}\right)^k \|\mathbf{e}^0\|_{\mathsf{A}}, \tag{5.15}$$

where $\kappa(\mathsf{A})$ denotes the spectral condition number of A.

Proof See, e.g., Axelsson [1] or Dostál [2]. □

The estimate (5.15) can be improved for some special distributions of the eigenvalues. For example, if $\sigma(\mathsf{A})$ is in a positive interval $[a_{\min}, a_{\max}]$ except for m isolated eigenvalues $\lambda_1, \dots, \lambda_m$, then we can use special polynomials $p \in \mathscr{P}^{k+m}$ of the form

$$p(\lambda) = \left(1 - \frac{\lambda}{\lambda_1}\right) \dots \left(1 - \frac{\lambda}{\lambda_m}\right) q(\lambda),\ q \in \mathscr{P}^k$$

to get the estimate

$$\|\mathbf{e}^{k+m}\|_{\mathsf{A}} \le 2\left(\frac{\sqrt{\widetilde{\kappa}} - 1}{\sqrt{\widetilde{\kappa}} + 1}\right)^k \|\mathbf{e}^0\|_{\mathsf{A}}, \tag{5.16}$$

where $\widetilde{\kappa} = a_{\max}/a_{\min}$.

If the spectrum of A satisfies $\sigma(\mathsf{A}) \subseteq [a_{\min}, a_{\max}] \cup [a_{\min} + d, a_{\max} + d]$, $d > 0$, then

$$\|\mathbf{e}^k\|_{\mathsf{A}} \leq 2\left(\frac{\sqrt{\overline{\kappa}} - 1}{\sqrt{\overline{\kappa}} + 1}\right)^k \|\mathbf{e}^0\|_{\mathsf{A}}, \tag{5.17}$$

where $\overline{\kappa} = 4a_{\max}/a_{\min}$ approximates the *effective condition number* of A. The proofs of the above bounds can be found in Axelsson [3] and Axelsson and Lindskøg [4].

5.4 Preconditioned Conjugate Gradients

The analysis of the previous section shows that the rate of convergence of the CG algorithm depends on the distribution of the eigenvalues of the Hessian A of f. In particular, we argued that CG converges very rapidly if the eigenvalues of A are clustered around one point, i.e., if the condition number $\kappa(\mathsf{A})$ is close to one. We shall now show that we can reduce our minimization problem to this favorable case if we have a symmetric positive definite matrix M such that $\mathsf{M}^{-1}\mathbf{x}$ can be easily evaluated for any $\mathbf{x}$ and M approximates A in the sense that $\mathsf{M}^{-1}\mathsf{A}$ is close to the identity.

First assume that M is available in the form

$$\mathsf{M} = \widetilde{\mathsf{L}}\widetilde{\mathsf{L}}^T,$$

so that $\mathsf{M}^{-1}\mathsf{A}$ is similar to $\widetilde{\mathsf{L}}^{-1}\mathsf{A}\widetilde{\mathsf{L}}^{-T}$ and the latter matrix is close to the identity. Then

$$f(\mathbf{x}) = \frac{1}{2}(\widetilde{\mathsf{L}}^T\mathbf{x})^T(\widetilde{\mathsf{L}}^{-1}\mathsf{A}\widetilde{\mathsf{L}}^{-T})(\widetilde{\mathsf{L}}^T\mathbf{x}) - (\widetilde{\mathsf{L}}^{-1}\mathbf{b})^T(\widetilde{\mathsf{L}}^T\mathbf{x})$$

and we can replace our original problem (5.1) by the *preconditioned problem* to find

$$\min_{\mathbf{y}\in\mathbb{R}^n} \bar{f}(\mathbf{y}), \tag{5.18}$$

where we substituted $\mathbf{y} = \widetilde{\mathsf{L}}^T\mathbf{x}$ and set

$$\bar{f}(\mathbf{y}) = \frac{1}{2}\mathbf{y}^T(\widetilde{\mathsf{L}}^{-1}\mathsf{A}\widetilde{\mathsf{L}}^{-T})\mathbf{y} - (\widetilde{\mathsf{L}}^{-1}\mathbf{b})^T\mathbf{y}.$$

The solution $\widehat{\mathbf{y}}$ of the preconditioned problem (5.18) is related to the solution $\widehat{\mathbf{x}}$ of the original problem by

$$\widehat{\mathbf{x}} = \widetilde{\mathsf{L}}^{-T}\widehat{\mathbf{y}}.$$

If the CG algorithm is applied directly to the preconditioned problem (5.18) with a given $\mathbf{y}^0$, then the algorithm is initialized by

$$\mathbf{y}^0 = \widetilde{\mathsf{L}}^T\mathbf{x}^0, \quad \bar{\mathbf{g}}^0 = \widetilde{\mathsf{L}}^{-1}\mathsf{A}\widetilde{\mathsf{L}}^{-T}\mathbf{y}^0 - \widetilde{\mathsf{L}}^{-1}\mathbf{b} = \widetilde{\mathsf{L}}^{-1}\mathbf{g}^0, \quad \text{and} \quad \bar{\mathbf{p}}^1 = \bar{\mathbf{g}}^0;$$

the iterates are defined by

$$\begin{aligned}
\bar{\alpha}_k &= \|\bar{\mathbf{g}}^{k-1}\|^2/(\bar{\mathbf{p}}^k)^T\widetilde{\mathsf{L}}^{-1}\mathsf{A}\widetilde{\mathsf{L}}^{-T}\bar{\mathbf{p}}^k,\\
\mathbf{y}^k &= \mathbf{y}^{k-1} - \bar{\alpha}_k\bar{\mathbf{p}}^k,\\
\bar{\mathbf{g}}^k &= \bar{\mathbf{g}}^{k-1} - \bar{\alpha}_k\widetilde{\mathsf{L}}^{-1}\mathsf{A}\widetilde{\mathsf{L}}^{-T}\bar{\mathbf{p}}^k,\\
\bar{\beta}_k &= \|\bar{\mathbf{g}}^k\|^2/\|\bar{\mathbf{g}}^{k-1}\|^2,\\
\bar{\mathbf{p}}^{k+1} &= \bar{\mathbf{g}}^k + \bar{\beta}_k\bar{\mathbf{p}}^k.
\end{aligned}$$

Substituting

$$\mathbf{y}^k = \widetilde{\mathsf{L}}^T\mathbf{x}^k, \quad \bar{\mathbf{g}}^k = \widetilde{\mathsf{L}}^{-1}\mathbf{g}^k, \quad \text{and} \quad \bar{\mathbf{p}}^k = \widetilde{\mathsf{L}}^T\mathbf{p}^k,$$

and denoting

$$\mathbf{z}^k = \widetilde{\mathsf{L}}^{-T}\widetilde{\mathsf{L}}^{-1}\mathbf{g}^k = \mathsf{M}^{-1}\mathbf{g}^k,$$

we obtain the *preconditioned conjugate gradient algorithm* (PCG) in the original variables.

Algorithm 5.2 Preconditioned conjugate gradient method (PCG).

Given an SPD matrix $\mathsf{A} \in \mathbb{R}^{n\times n}$, *its SPD approximation* $\mathsf{M} \in \mathbb{R}^{n\times n}$, *and* $\mathbf{b} \in \mathbb{R}^n$.

Step 0. {*Initialization.*}

Choose $\mathbf{x}^0 \in \mathbb{R}^n$, *set* $\mathbf{g}^0 = \mathsf{A}\mathbf{x}^0 - \mathbf{b}$, $\mathbf{z}^0 = \mathsf{M}^{-1}\mathbf{g}^0$, $\mathbf{p}^1 = \mathbf{z}^0$, $k = 1$

Step 1. {*Conjugate gradient loop.*}

while $\|\mathbf{g}^{k-1}\|$ *is not small*

$\alpha_k = (\mathbf{z}^{k-1})^T\mathbf{g}^{k-1}/(\mathbf{p}^k)^T\mathsf{A}\mathbf{p}^k$

$\mathbf{x}^k = \mathbf{x}^{k-1} - \alpha_k\mathbf{p}^k$

$\mathbf{g}^k = \mathbf{g}^{k-1} - \alpha_k\mathsf{A}\mathbf{p}^k$

$\mathbf{z}^k = \mathsf{M}^{-1}\mathbf{g}^k$

$\beta_k = (\mathbf{z}^k)^T\mathbf{g}^k/(\mathbf{z}^{k-1})^T\mathbf{g}^{k-1}$

$\mathbf{p}^{k+1} = \mathbf{z}^k + \beta_k\mathbf{p}^k$

$k = k + 1$

end while

Step 2. {*Return a (possibly approximate) solution.*}

$\widetilde{\mathbf{x}} = \mathbf{x}^k$

Notice that the PCG algorithm does not explicitly exploit the Cholesky factorization of the preconditioner M. The *pseudoresiduals* $\mathbf{z}^k$ are typically obtained by solving $\mathsf{M}\mathbf{z}^k = \mathbf{g}^k$. If M is a good approximation of A, then $\mathbf{z}^k$ is close to the error vector $\mathbf{e}^k$. The rate of convergence of the PCG algorithm depends on the condition number of the Hessian of the transformed function $\tilde{f}$, i.e., on $\kappa(\mathsf{M}^{-1}\mathsf{A}) = \kappa(\tilde{\mathsf{L}}^{-1}\mathsf{A}\tilde{\mathsf{L}}^{-T})$. Thus the efficiency of the preconditioned CG method depends critically on the choice of a preconditioner, which should balance the cost of its application with the preconditioning effect. We refer interested readers to specialized books like Axelsson [1] or Saad [5] for more information. In our applications, we use the standard FETI preconditioners introduced by Fahat, Mandel, and Roux [6].

5.5 Convergence in Presence of Rounding Errors

The elegant mathematical theory presented above assumes implementation of the CG algorithm in exact arithmetic and captures well the performance of only a limited number of CG iterations in computer arithmetics. Since we use the CG method mainly for a low-precision approximation of well-conditioned auxiliary problems, we shall base our exposition on this theory in what follows. However, it is still useful to be aware of possible effects of rounding errors that accompany any computer implementation of the CG algorithm, especially for the solution of very large problems.

It has been known since the introduction of the CG method that, when used in finite precision arithmetic, the vectors generated by these algorithms can seriously violate their theoretical properties. In particular, it has been observed that the evaluated gradients can lose their orthogonality after as small a number of iterations as twenty, and that nearly dependent conjugate directions can be generated. In spite of these effects, it has been observed that the CG method still converges in finite precision arithmetic, but that the convergence is delayed [7, 8].

Undesirable effects of the rounding errors can be reduced by reorthogonalization. A simple analysis reveals that the full reorthogonalization of the gradients is costly and requires large memory. A key to an efficient implementation of the reorthogonalization is based on observation that accumulation of the rounding errors has a regular pattern, namely, that large perturbations of the generated vectors belong to the space generated by the eigenvectors of A which can be approximated well by the vectors from the current Krylov space. This has led to the efficient implementation of the CG method based on the *selective orthogonalization* proposed by Parlett and Scott [9]. More details and information about the effects of rounding errors and implementation of the CG method in finite arithmetic can be found in the comprehensive review paper by Meurant and Strakoš [10].

5.6 Comments and Conclusions

Since its introduction in the early 1950s by Hestenes and Stiefel [11], a lot of research related to the development of the CG method has been carried out, so that there are many references concerning this subject. We refer an interested reader to the textbooks and research monographs by Saad [5], van der Vorst [12], Greenbaum [13], and Axelsson [1] for more information. A comprehensive account of development of the CG method up to 1989 may be found in the paper by Golub and O'Leary [14]. Most of the research is concentrated on the development and analysis of preconditioners.

Finding at each step the minimum over the subspace generated by all the previous search directions, the CG method exploits all information gathered during the previous iterations. To use this feature in the algorithms for the solution of constrained problems, it is important to generate long uninterrupted sequences of the CG iterations. This strategy also supports exploitation of yet another unique feature of the CG method, namely, its self-preconditioning capabilities that were described by van der Sluis and van der Vorst [15]. The latter property can also be described in terms of the preconditioning by the conjugate projector. Indeed, if Q_k denotes the conjugate projector onto the conjugate complement V of $U = \mathrm{Span}\{\mathbf{p}_1, \ldots, \mathbf{p}_k\}$, then it is possible to give the bound on the rate of convergence of the CG method starting from $\mathbf{x}_{k+1}$ in terms of the spectral *regular condition number* $\overline{\kappa}_k = \overline{\kappa}(\mathsf{Q}_k^T \mathsf{A} \mathsf{Q}_k | V)$ of $\mathsf{Q}_k^T \mathsf{A} \mathsf{Q}_k | V$ and observe that $\overline{\kappa}_k$ decreases with the increasing k. Recall that the spectral regular condition number $\overline{\kappa}(\mathsf{A})$ of an SPS matrix A is defined by

$$\overline{\kappa}(\mathsf{A}) = \frac{\|\mathsf{A}\|}{\overline{\lambda}_{\min}(\mathsf{A})},$$

where $\overline{\lambda}_{\min}(\mathsf{A})$ denotes the least nonzero eigenvalue of A.

References

1. Axelsson, O.: Iterative Solution Methods. Cambridge University Press, Cambridge (1994)
2. Dostál, Z.: Optimal Quadratic Programming Algorithms, with Applications to Variational Inequalities, 1st edn. Springer, New York (2009)
3. Axelsson, O.: A class of iterative methods for finite element equations. Comput. Methods Appl. Mech. Eng. **9**, 127–137 (1976)
4. Axelsson, O., Lindskøg, G.: On the rate of convergence of the preconditioned conjugate gradient method. Numer. Math. **48**, 499–523 (1986)
5. Saad, Y.: Iterative Methods for Large Linear Systems. SIAM, Philadelphia (2002)
6. Farhat, C., Mandel, J., Roux, F.-X.: Optimal convergence properties of the FETI domain decomposition method. Comput. Methods Appl. Mech. Eng. **115**, 365–385 (1994)
7. Greenbaum, A.: Behaviour of slightly perturbed Lanczos and conjugate gradient algorithms. Linear Algebr. Appl. **113**, 7–63 (1989)
8. Greenbaum, A., Strakoš, Z.: Behaviour of slightly perturbed Lanczos and conjugate gradient algorithms. SIAM J. Matrix Anal. Appl. **22**, 121–137 (1992)

9. Parlett, B.N., Scott, D.S.: The Lanczos algorithm with selective orthogonalization. Math. Comput. **33**, 217–238 (1979)
10. Meurant, G., Strakoš, Z.: The Lanczos and conjugate gradient algorithms in finite precision arithmetic. Acta Numer. 471–542 (2006)
11. Hestenes, M.R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Res. National Bureau Stand. **49**, 409–436 (1952)
12. van der Vorst, H.: Iterative Krylov Methods for Large Linear Systems. Cambridge University Press, Cambridge (2003)
13. Greenbaum, A.: Iterative Methods for Solving Linear Systems. SIAM, Philaledelphia (1997)
14. Golub, G.H., O'Leary, D.P.: Some history of the conjugate gradient and Lanczos methods. SIAM Review **31**, 50–102 (1989)
15. van der Sluis, A., van der Vorst, H.A.: The rate of convergence of the conjugate gradients. Numer. Math. **48**, 543–560 (1986)

Chapter 6
Gradient Projection for Separable Convex Sets

An important ingredient of our algorithms for solving QP and QCQP problems is the Euclidean projection on the convex set defined by separable convex constraints. To combine the gradient projection with the CG method effectively, it is necessary to have nontrivial bounds on the decrease of f along the projected-gradient path in terms of bounds on the spectrum of its Hessian matrix A. While such results are standard for the solution of unconstrained quadratic programming problems [1, 2], it seems that until recently there were no such results for inequality constrained problems. The standard results either provide the bounds on the contraction of the gradient projection [3] in the Euclidean norm or guarantee only some qualitative properties of convergence (see, e.g., Luo and Tseng [4]). Here we present the results concerning the decrease of f along the projected-gradient path in the extent that is necessary for the development of scalable algorithms for contact problems.

6.1 Separable Convex Constraints and Projections

Our goal is to get insight into the effect of the projected-gradient step for the problem to find

$$\min_{\mathbf{x}\in\Omega} f(\mathbf{x}), \tag{6.1}$$

where $f = \frac{1}{2}\mathbf{x}^T\mathsf{A}\mathbf{x} - \mathbf{x}^T\mathbf{b}$, $\mathsf{A} \in \mathbb{R}^{n\times n}$ denotes an SPS matrix, $\mathbf{b}, \mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{x} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_s^T]^T, \quad \mathbf{x}_i = [\mathbf{x}_1^i, \ldots, \mathbf{x}_{\ell_i}^i], \quad \ell_1 + \cdots + \ell_s = n,$$

and

$$\Omega = \Omega_1 \times \cdots \times \Omega_s$$

denotes a closed convex set defined by separable constraints $h_i : \mathbb{R}^{\ell_i} \to \mathbb{R}$, $i = 1, \ldots, s$, i.e.,

$$\Omega_i = \{x_i \in \mathbb{R}^{\ell_i} : h_i(x_i) \le 0\},\ i = 1, \ldots, s.$$

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3_6

We assume that $\mathbf{b}$ has the same block structure as $\mathbf{x}$, i.e.,

$$\mathbf{b} = [\mathbf{b}_1^T, \ldots, \mathbf{b}_s^T]^T, \quad \mathbf{b}_i \in \mathbb{R}^{\ell_i}.$$

We are especially interested in the problems defined by the separable elliptic constraints

$$h_i(\mathbf{x}_i) = (\mathbf{x}_i - \mathbf{y}_i)^T \mathsf{H}_i (\mathbf{x}_i - \mathbf{y}_i) - d_i, \quad \mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^2, \quad d_i > 0, \quad \mathsf{H}_i \text{ SPD}, \tag{6.2}$$

that appear in the dual formulation of contact problems with Tresca friction, or the bound constraints

$$h_i(x_i) = \ell_i - x_i, \quad \ell_i, x_i \in \mathbb{R},$$

arising in the dual formulation of non-penetration conditions. In what follows, we denote by P_Ω the Euclidean projection to Ω, so that

$$P_\Omega(\mathbf{x}) = \arg\min_{\mathbf{y}\in\Omega} \|\mathbf{x} - \mathbf{y}\|.$$

Since the constraints that define Ω are separable, we can define P_Ω block-wise by

$$P_{\Omega_i}(\mathbf{x}_i) = \arg\min_{\mathbf{y}\in\Omega_i} \|\mathbf{x}_i - \mathbf{y}\|, \quad P_\Omega(\mathbf{x}) = \left[P_{\Omega_1}(\mathbf{x}_1)^T, \ldots, P_{\Omega_s}(\mathbf{x}_s)^T\right]^T. \tag{6.3}$$

The action of P_Ω is especially easy to calculate for the spherical or bound constraints. As illustrated by Fig. 6.1, the components of the projection $P_{\Omega_B}(\mathbf{x})$ of $\mathbf{x}$ onto

$$\Omega_B = \{\mathbf{x} \in \mathbb{R}^n : x_i \geq \ell_i, \ i = 1, \ldots, n\}$$

are given by

$$[P_{\Omega_B}(\mathbf{x})]_i = \max\{\ell_i, x_i\}, \ i = 1, \ldots, n.$$

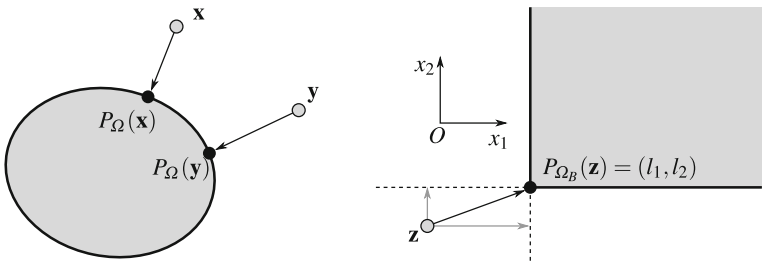

Fig. 6.1 Euclidean projections onto convex sets

The gradient projection is an important ingredient of the minimization algorithms. A typical step of the gradient projection method is in Fig. 6.2.

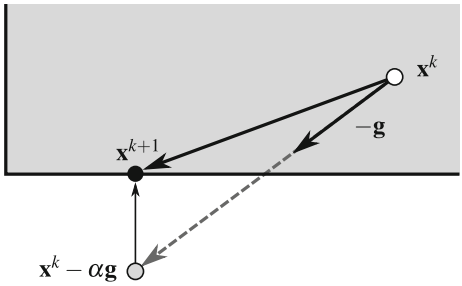

Fig. 6.2 Gradient projection step

6.2 Conjugate Gradient Step Versus Gradient Projections

Since the conjugate gradient is the best decrease direction which can be used to find the minimizer in a current Krylov space by Theorem 5.1, probably the first idea how to plug the projection into CG-based algorithms for (6.1) is to replace the conjugate gradient step by the projected conjugate gradient step

$$\mathbf{x}^{k+1} = P_\Omega(\mathbf{x}^k - \alpha_{cg}\mathbf{p}^k).$$

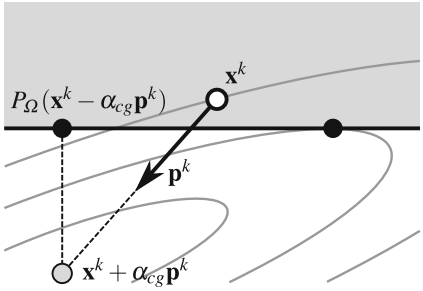

Fig. 6.3 Poor performance of the projected conjugate gradient step

However, if we examine Fig. 6.3, which depicts the 2D situation after the first conjugate gradient step for a bound constrained problem, we can see that though the second conjugate gradient step finds the unconstrained minimizer $\mathbf{x}^k - \alpha_{cg}\mathbf{p}^k$, it can easily happen that

$$f(\mathbf{x}^k) < f(P_\Omega(\mathbf{x}^k - \alpha_{cg}\mathbf{p}^k)).$$

Figure 6.3 even suggests that it is possible that for any $\alpha > \alpha_f$

$$f(P_\Omega(\mathbf{x}^k - \alpha\mathbf{p}^k)) > f\big(P_\Omega(\mathbf{x}^k - \alpha_f\mathbf{p}^k)\big).$$

Though Fig. 6.3 need not capture the typical situation when a small number of components of $\mathbf{x}^k - \alpha_f \mathbf{p}^k$ is affected by P_Ω, we conclude that the nice properties of conjugate directions are guaranteed only in the feasible region. These observations comply with our discussion at the end of Sect. 5.3.

On the other hand, since the gradient defines the direction of the steepest descent, it is natural to assume that for a small step length the gradient perturbed by the projection P_Ω defines a decrease direction as in Fig. 6.4. We shall prove a quantitative refinement of this conjecture. In what follows, we restrict our attention to the analysis of the fixed steplength gradient iteration

$$\mathbf{x}^{k+1} = P_\Omega(\mathbf{x}^k - \alpha \mathbf{g}^k), \tag{6.4}$$

where $\mathbf{g}^k = \nabla f(\mathbf{x}^k)$.

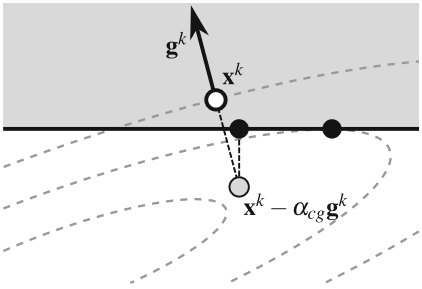

Fig. 6.4 Fixed steplength gradient step

Which values of α guarantee that the iterates defined by the fixed gradient projection step (6.4) approach the solution $\widehat{\mathbf{x}}$ in the Euclidean norm?

Proposition 6.1 *Let Ω denote a closed convex set, $\mathbf{x} \in \Omega$, and $\mathbf{g} = \nabla f(\mathbf{x})$. Then for any $\alpha > 0$*

$$\| P_\Omega(\mathbf{x} - \alpha \mathbf{g}) - \widehat{\mathbf{x}} \| \leq \eta_E \| \mathbf{x} - \widehat{\mathbf{x}} \|, \tag{6.5}$$

where $\lambda_{\min}$, $\lambda_{\max}$ are the extreme eigenvalues of A and

$$\eta_E = \max\{|1 - \alpha\lambda_{\min}|, |1 - \alpha\lambda_{\max}|\}. \tag{6.6}$$

Proof Since $\widehat{\mathbf{x}} \in \Omega$ and the projected-gradient at the solution satisfies $\widehat{\mathbf{g}}^P = \mathbf{o}$, it follows that

$$P_\Omega(\widehat{\mathbf{x}} - \alpha\widehat{\mathbf{g}}) = \widehat{\mathbf{x}}.$$

Using that the projection P_Ω is nonexpansive by Corollary 3.1, the definition of gradient, the relations between the norm of a symmetric matrix and its spectrum (2.23), and the observation that if λ_i are the eigenvalues of A, then $1 - \alpha\lambda_i$ are the eigenvalues of $\mathsf{I} - \alpha\mathsf{A}$ (see also (2.25)), and we get

$$\begin{aligned}\|P_\Omega(\mathbf{x}-\alpha\mathbf{g})-\widehat{\mathbf{x}}\| &= \|P_\Omega(\mathbf{x}-\alpha\mathbf{g})-P_\Omega(\widehat{\mathbf{x}}-\alpha\widehat{\mathbf{g}})\| \\ &\le \|(\mathbf{x}-\alpha\mathbf{g})-(\widehat{\mathbf{x}}-\alpha\widehat{\mathbf{g}})\| \\ &= \|\,(\mathbf{x}-\widehat{\mathbf{x}})-\alpha(\mathbf{g}-\widehat{\mathbf{g}})\| = \|\,(\mathbf{x}-\widehat{\mathbf{x}})-\alpha\mathsf{A}(\mathbf{x}-\widehat{\mathbf{x}})\| \\ &= \|(\mathsf{I}-\alpha\mathsf{A})(\mathbf{x}-\widehat{\mathbf{x}})\| \\ &\le \max\{|1-\alpha\lambda_{\min}|, |1-\alpha\lambda_{\max}|\}\|\mathbf{x}-\widehat{\mathbf{x}}\|. \end{aligned}$$

□

We call η_E the *coefficient of Euclidean contraction*. If $\alpha \in (0, 2\|\mathsf{A}\|^{-1})$, then $\eta_E < 1$. Using some elementary arguments, we get that η_E is minimized by

$$\alpha_E^{opt} = \frac{2}{\lambda_{\min}+\lambda_{\max}} \tag{6.7}$$

and

$$\eta_E^{opt} = \frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}} = \frac{\kappa-1}{\kappa+1}, \quad \kappa = \frac{\lambda_{\max}}{\lambda_{\min}}. \tag{6.8}$$

Notice that the estimate (6.5) does not guarantee any bound on the decrease of the cost function. We study this topic in Sect. 6.5.

6.3 Quadratic Functions with Identity Hessian

Which values of α guarantee that the cost function f decreases in each iterate defined by the fixed gradient projection step (6.4)? How much does f decrease when the answer is positive? To answer these questions, it is useful to carry out some analysis for a special quadratic function

$$F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{x} - \mathbf{c}^T\mathbf{x}, \quad \mathbf{x}\in\mathbb{R}^n, \tag{6.9}$$

which is defined by a fixed $\mathbf{c}\in\mathbb{R}^n$, $\mathbf{c}=[c_i]$. We shall also use

$$F(\mathbf{x}) = \sum_{i=1}^n F_i(x_i), \quad F_i(x_i) = \frac{1}{2}x_i^2 - c_i x_i, \quad \mathbf{x}=[x_i]. \tag{6.10}$$

The Hessian and the gradient of F are expressed by

$$\nabla^2 F(\mathbf{x}) = \mathsf{I} \quad \text{and} \quad \mathbf{g} = \nabla F(\mathbf{x}) = \mathbf{x}-\mathbf{c}, \quad \mathbf{g}=[g_i], \tag{6.11}$$

respectively. Thus $\mathbf{c}=\mathbf{x}-\mathbf{g}$ and for any $\mathbf{z}\in\mathbb{R}^n$

$$\|\mathbf{z}-\mathbf{c}\|^2 = \|\mathbf{z}\|^2 - 2\mathbf{c}^T\mathbf{z} + \|\mathbf{c}\|^2 = 2F(\mathbf{z}) + \|\mathbf{c}\|^2.$$

Since by Proposition 3.5 for any $\mathbf{z}\in\Omega$

$$\|\mathbf{z}-\mathbf{c}\| \geq \|P_{\Omega}(\mathbf{c})-\mathbf{c}\|,$$

we get that for any $\mathbf{z} \in \Omega$

$$\begin{aligned} 2F(\mathbf{z}) &= \|\mathbf{z}-\mathbf{c}\|^2 - \|\mathbf{c}\|^2 \geq \|P_{\Omega}(\mathbf{c})-\mathbf{c}\|^2 - \|\mathbf{c}\|^2 \\ &= 2F\left(P_{\Omega}(\mathbf{c})\right) = 2F\left(P_{\Omega}(\mathbf{x}-\mathbf{g})\right). \end{aligned} \tag{6.12}$$

We have thus proved that if $\mathbf{y} \in \Omega$, then, as illustrated in Fig. 6.5,

$$F\big(P_{\Omega}(\mathbf{x}-\mathbf{g})\big) \leq F(\mathbf{y}). \tag{6.13}$$

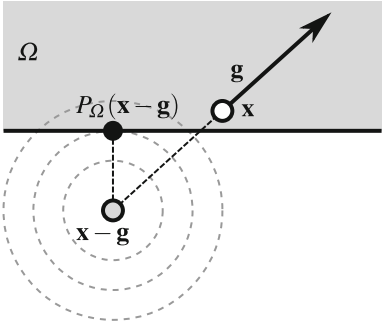

Fig. 6.5 Minimizer of F in Ω

We are especially interested in the analysis of F along the projected-gradient path

$$p(\mathbf{x},\alpha) = P_{\Omega}\big(\mathbf{x}-\alpha\nabla F(\mathbf{x})\big),$$

where $\alpha \geq 0$ and $\mathbf{x} \in \Omega$ is fixed. A geometric illustration of the projected-gradient path for the feasible set of a bound constrained problem is in Fig. 6.6.

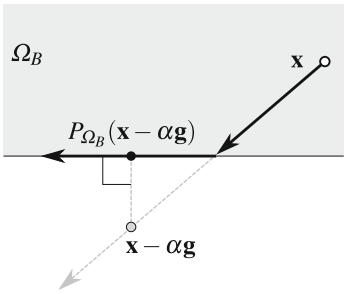

Fig. 6.6 Projected-gradient path

6.4 Subsymmetric Sets

In the analysis of the rate of convergence in Chaps. 7 and 8, we shall use the following generalization of a property of the half-interval (see Dostál [5]).

Definition 6.1 A closed convex set $\Omega \subseteq \mathbb{R}^n$ is *subsymmetric* if for any $\mathbf{x} \in \Omega$, $\delta \in [0, 1]$, $\mathbf{c} \in \mathbb{R}^n$, and $\mathbf{g} = \mathbf{x} - \mathbf{c}$

$$F\big(P_\Omega(\mathbf{x} - (2-\delta)\mathbf{g})\big) \le F\big(P_\Omega(\mathbf{x} - \delta\mathbf{g})\big), \tag{6.14}$$

where F is defined by (6.9).

The condition which defines the subsymmetric set is illustrated in Fig. 6.7.

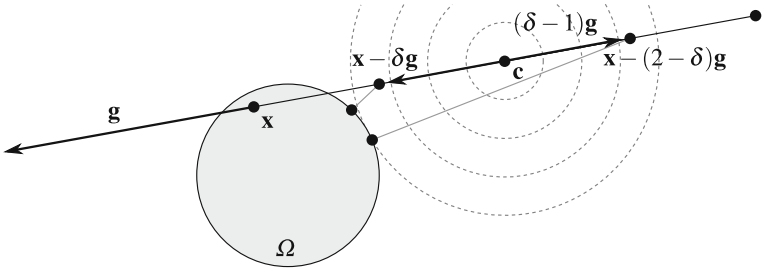

Fig. 6.7 The condition which defines a subsymmetric set

Let us show that the half-interval is subsymmetric.

Lemma 6.1 *Let $\ell \in \mathbb{R}$ and $\Omega_B = [\ell, \infty)$. Let F and g be defined by*

$$F(x) = \frac{1}{2}x^2 - cx \quad \text{and} \quad g = x - c.$$

Then for any $\delta \in [0, 1]$

$$F\big(P_{\Omega_B}(x - (2-\delta)g)\big) \le F\big(P_{\Omega_B}(x - \delta g)\big). \tag{6.15}$$

Proof First assume that $x \ge l$ is fixed and denote

$$g = F'(x) = x - c, \quad \widetilde{g}(0) = 0, \quad \widetilde{g}(\alpha) = \min\{(x - \ell)/\alpha, g\}, \quad \alpha \ne 0.$$

For convenience, let us define

$$F\big(P_{\Omega_B}(x - \alpha g)\big) = F(x) + \Phi(\alpha), \quad \Phi(\alpha) = -\alpha\widetilde{g}(\alpha)g + \frac{\alpha^2}{2}\left(\widetilde{g}(\alpha)\right)^2, \quad \alpha \ge 0.$$

Moreover, using these definitions, it can be checked directly that Φ is defined by

$$\Phi(\alpha) = \begin{cases} \Phi_F(\alpha) & \text{for} \quad \alpha \in (-\infty, \overline{\xi}\,] \cap [0, \infty) \quad \text{or} \quad g \le 0, \\ \Phi_A(\alpha) & \text{for} \quad \alpha \in [\ \overline{\xi}\ , \infty) \cap [0, \infty) \quad \text{and} \quad g > 0, \end{cases}$$

where $\overline{\xi} = \infty$ if $g = 0$ and $\overline{\xi} = (x - \ell)/g$ if $g \neq 0$,

$$\Phi_F(\alpha) = \left(-\alpha + \frac{\alpha^2}{2}\right) g^2, \quad \text{and} \quad \Phi_A(\alpha) = -g(x - \ell) + \frac{1}{2}(x - \ell)^2.$$

See also Fig. 6.8.

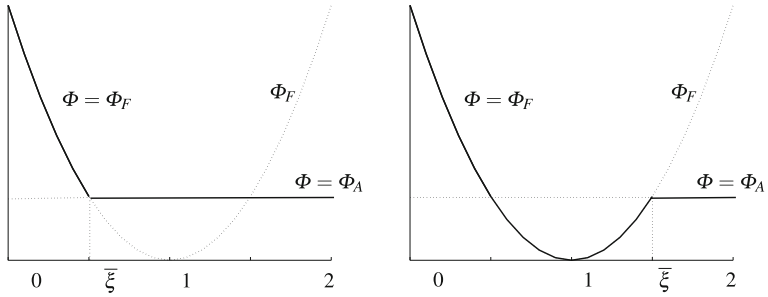

Fig. 6.8 Graphs of Φ for $\overline{\xi} < 1$ (*left*) and $\overline{\xi} > 1$ (*right*) when $g > 0$

It follows that for any α

$$\Phi_F(2 - \alpha) = \left(-(2 - \alpha) + \frac{(2 - \alpha)^2}{2}\right) g^2 = \Phi_F(\alpha), \tag{6.16}$$

and if $g \le 0$, then

$$\Phi(\alpha) = \Phi_F(\alpha) = \Phi_F(2 - \alpha) = \Phi(2 - \alpha).$$

Let us now assume that $g > 0$ and denote $\overline{\xi} = (x - \ell)/g$. Simple analysis shows that if $\overline{\xi} \in [0, 1]$, then Φ is nonincreasing on $[0, 2]$ and (6.15) is satisfied for $\alpha \in [0, 1]$. To finish the proof of (6.15), notice that if $1 < \overline{\xi}$, then

$$\Phi(\alpha) = \Phi_F(\alpha), \quad \alpha \in [0, 1], \quad \Phi(\alpha) \le \Phi_F(\alpha), \quad \alpha \in [1, 2],$$

so that we can use (6.16) to get that for $\alpha \in [0, 1]$

$$\Phi(2 - \alpha) \le \Phi_F(2 - \alpha) = \Phi_F(\alpha) = \Phi(\alpha).$$

□

Bouchala and Vodstrčil managed to prove that also the ellipse is a subsymmetric set [6].

Lemma 6.2 *Let $\Omega \subset \mathbb{R}^2$ be defined by an elliptic constraint*

$$\Omega = \{\mathbf{x} \in \mathbb{R}^2 : h(\mathbf{x}) \le 0\}, \quad h(\mathbf{x}) = (\mathbf{x}-\mathbf{y})^T \mathsf{H}(\mathbf{x}-\mathbf{y}) - d, \quad \mathbf{y} \in \mathbb{R}^2, \quad d > 0,$$

where H is SPD. Then Ω is subsymmetric.

Proof Clearly, we can assume that the ellipse Ω is centered at the origin, i.e.,

$$\Omega = \left\{ \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2 : \frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} \le 1 \right\},$$

where $a, b \in \mathbb{R}$ and $a, b > 0$. We prove the following three cases:

1. $\mathbf{x} + \alpha\mathbf{g} \in \Omega,\ \mathbf{x} - \alpha\mathbf{g} \in \Omega,$
2. $\mathbf{x} + \alpha\mathbf{g} \in \Omega,\ \mathbf{x} - \alpha\mathbf{g} \notin \Omega,$
3. $\mathbf{x} + \alpha\mathbf{g} \notin \Omega,\ \mathbf{x} - \alpha\mathbf{g} \notin \Omega.$

Ad 1. In this case,

$$\|P_\Omega(\mathbf{x} + \alpha\mathbf{g}) - \mathbf{x}\| = \|\alpha\mathbf{g}\| = \|P_\Omega(\mathbf{x} - \alpha\mathbf{g}) - \mathbf{x}\|.$$

Ad 2. Since $[b\cos t, a\sin t]$ is the outer normal vector of the ellipse

$$\partial\Omega = \{[a\cos t, b\sin t] \in \mathbb{R}^2 :\ t \in \mathbb{R}\}$$

at $[a\cos t, b\sin t]$, we can write $\mathbf{x} + \alpha\mathbf{g}$ and $\mathbf{x} - \alpha\mathbf{g}$ in the form

$$\begin{aligned} \mathbf{x} + \alpha\mathbf{g} &= s[a\cos t_1, b\sin t_1]^T, \\ \mathbf{x} - \alpha\mathbf{g} &= [a\cos t_2, b\sin t_2]^T + k[b\cos t_2, a\sin t_2]^T, \end{aligned} \tag{6.17}$$

where $0 \le s \le 1,\ k > 0,\ t_1, t_2 \in \mathbb{R}$. It can be verified directly that

$$\|P_\Omega(\mathbf{x} + \alpha g) - \mathbf{x}\|^2 - \|P_\Omega(\mathbf{x} - \alpha g) - \mathbf{x}\|^2 = abk\big(1 - s\cos(t_2 - t_1)\big) \ge 0.$$

Ad 3. Now we can write

$$\begin{aligned} \mathbf{x} + \alpha\mathbf{g} &= [a\cos t_1, b\sin t_1]^T + s[b\cos t_1, a\sin t_1]^T, \\ \mathbf{x} - \alpha\mathbf{g} &= [a\cos t_2, b\sin t_2]^T + k[b\cos t_2, a\sin t_2]^T, \end{aligned} \tag{6.18}$$

where $s, k > 0$, and we can check that

$$\|P_\Omega(\mathbf{x} + \alpha\mathbf{g}) - \mathbf{x}\|^2 - \|P_\Omega(\mathbf{x} - \alpha\mathbf{g}) - \mathbf{x}\|^2 = ab(k - s)\big(1 - \cos(t_1 - t_2)\big).$$

It remains to prove that $k \ge s$. Due to the symmetry of the ellipse, we can assume $0 < t_2 \le t_1 < \pi$. Now we consider such $t_0 \in (t_1, \pi]$ that the line given by two points

$$\mathbf{x} + \alpha\mathbf{g} = [a\cos t_1, b\sin t_1]^T + s[b\cos t_1, a\sin t_1]^T \quad \text{and} \quad [a\cos t_0, b\sin t_0]^T$$

is a tangent to the ellipse. It is easy to calculate that

$$s = \frac{ab\big(1-\cos(t_0-t_1)\big)}{b^2\cos t_0\cos t_1 + a^2\sin t_0\sin t_1}, \quad k \geq \frac{ab\big(1-\cos(t_0-t_2)\big)}{b^2\cos t_0\cos t_2 + a^2\sin t_0\sin t_2},$$

and that the function

$$f(t) := \frac{ab\big(1-\cos(t_0-t)\big)}{b^2\cos t_0\cos t + a^2\sin t_0\sin t}$$

has a nonpositive derivative (and therefore f is nonincreasing) on the interval $[t_2, t_1]$. This implies $k \geq s$. □

Now we are ready to prove the main result of this section.

Proposition 6.2 *Let $\Omega \subset \mathbb{R}^n$ be defined as a direct product of ellipses and/or halfspaces. Then Ω is subsymmetric, i.e.,*

$$F\big(P_\Omega(\mathbf{x}-(2-\delta)\mathbf{g})\big) \leq F\big(P_\Omega(\mathbf{x}-\delta\mathbf{g})\big). \tag{6.19}$$

Proof Let

$$\Omega = \Omega_1 \times \cdots \times \Omega_s, \tag{6.20}$$

where Ω_i is either a half-interval or an ellipse. If $s = 1$, then the statement reduces to Lemma 6.1 or Lemma 6.2.

To prove the statement for $s > 1$, first observe that for any $\mathbf{y} \in \mathbb{R}$

$$[P_\Omega(\mathbf{y})]_i = P_{\Omega_i}(\mathbf{y}_i),\ i = 1, \ldots, s.$$

It follows that P_Ω is separable and can be defined componentwise by the real functions

$$P_i(y) = \max\{y, \ell_i\},\ i = 1, \ldots, n.$$

Using the separable representation of F given by (6.10), we can define a separable representation of F in the form

$$F(\mathbf{x}) = \sum_{i=1}^{s} F_i(\mathbf{x}_i),\ i = 1, \ldots, s,$$

which complies with (6.20). To complete the proof, it is enough to use this representation, Lemma 6.1, and Lemma 6.2 to get

$$\begin{aligned} F\big(P_\Omega(\mathbf{x}-(2-\delta)\mathbf{g})\big) &= \sum_{i=1}^{s} F_i\big([P_\Omega(\mathbf{x}-(2-\delta)\mathbf{g})]_i\big) \\ &= \sum_{i=1}^{n} F_i\big(P_{\Omega_i}(\mathbf{x}_i-(2-\delta)\mathbf{g}_i)\big) \\ &\leq \sum_{i=1}^{n} F_i\big(P_{\Omega_i}(\mathbf{x}_i-\delta\mathbf{g}_i)\big) \\ &= F\big(P_\Omega(\mathbf{x}-\delta\mathbf{g})\big). \end{aligned}$$

□

6.5 Dominating Function and Decrease of the Cost Function

Now we are ready to give an estimate of the decrease of the cost function

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \mathsf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x}$$

along the projected-gradient path. The idea of the proof is to replace f by a suitable quadratic function F which dominates f and has the Hessian equal to the identity matrix.

Let us assume that Ω is convex, $0 < \delta\|\mathsf{A}\| \leq 1$, and let $\mathbf{x} \in \Omega$ be arbitrary but fixed, so that we can define a quadratic function

$$F_\delta(\mathbf{y}) = \delta f(\mathbf{y}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^T(\mathsf{I} - \delta\mathsf{A})(\mathbf{y}-\mathbf{x}), \quad \mathbf{y} \in \mathbb{R}^n. \tag{6.21}$$

It is defined so that

$$F_\delta(\mathbf{x}) = \delta f(\mathbf{x}), \quad \nabla F_\delta(\mathbf{x}) = \delta\nabla f(\mathbf{x}) = \delta\mathbf{g}, \quad \text{and} \quad \nabla^2 F_\delta(\mathbf{y}) = \mathsf{I}. \tag{6.22}$$

Moreover, for any $\mathbf{y} \in \mathbb{R}^n$

$$\delta f(\mathbf{y}) \leq F_\delta(\mathbf{y}). \tag{6.23}$$

It follows that

$$\delta f\left(P_\Omega(\mathbf{x} - \delta\mathbf{g})\right) - \delta f(\widehat{\mathbf{x}}) \leq F_\delta\left(P_\Omega(\mathbf{x} - \delta\mathbf{g})\right) - \delta f(\widehat{\mathbf{x}}) \tag{6.24}$$

and

$$\nabla F_\delta(\mathbf{y}) = \delta\nabla f(\mathbf{y}) + (\mathsf{I} - \delta\mathsf{A})(\mathbf{y}-\mathbf{x}) = \mathbf{y} - (\mathbf{x} - \delta\mathbf{g}). \tag{6.25}$$

Using (6.13) and (6.22), we get that for any $\mathbf{z} \in \Omega$

$$F_\delta\left(P_\Omega(\mathbf{x} - \delta\mathbf{g})\right) \leq F_\delta(\mathbf{z}). \tag{6.26}$$

The following lemma is due to Schöberl [7, 8].

Lemma 6.3 *Let Ω be a closed convex set, let $\lambda_{\min}$ denote the smallest eigenvalue of A, $\mathbf{g} = \nabla f(\mathbf{x})$, $\mathbf{x} \in \Omega$, $\delta \in (0, \|\mathsf{A}\|^{-1}]$, and let*

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\Omega} f(x)$$

denote a unique solution of (8.1). Then

$$F_\delta\left(P_\Omega(\mathbf{x} - \delta\mathbf{g})\right) - \delta f(\widehat{\mathbf{x}}) \leq \delta(1 - \delta\lambda_{\min})\left(f(\mathbf{x}) - f(\widehat{\mathbf{x}})\right). \tag{6.27}$$

Proof Let us denote

$$[\widehat{\mathbf{x}}, \mathbf{x}] = \mathrm{Conv}\{\widehat{\mathbf{x}}, \mathbf{x}\} \quad \text{and} \quad \mathbf{d} = \widehat{\mathbf{x}} - \mathbf{x}.$$

Using (6.26),

$$[\widehat{\mathbf{x}}, \mathbf{x}] = \{\mathbf{x} + t\mathbf{d} : t \in [0, 1]\} \subseteq \Omega,$$

$0 < \lambda_{\min}\delta \leq \|\mathsf{A}\|\delta \leq 1$, and $\lambda_{\min}\|\mathbf{d}\|^2 \leq \mathbf{d}^T\mathsf{A}\mathbf{d}$, we get

$$\begin{aligned}
F_\delta\left(P_\Omega(\mathbf{x} - \delta\mathbf{g})\right) - \delta f(\widehat{\mathbf{x}}) &= \min\{F_\delta(\mathbf{y}) - \delta f(\widehat{\mathbf{x}}) : \ \mathbf{y} \in \Omega\} \\
&\leq \min\{F_\delta(\mathbf{y}) - \delta f(\widehat{\mathbf{x}}) : \ \mathbf{y} \in [\widehat{\mathbf{x}}, \mathbf{x}]\} \\
&= \min\{F_\delta(\mathbf{x} + t\mathbf{d}) - \delta f(\mathbf{x} + \mathbf{d}) : \ t \in [0, 1]\} \\
&= \min\{\delta t\mathbf{d}^T\mathbf{g} + \frac{t^2}{2}\|\mathbf{d}\|^2 - \delta\mathbf{d}^T\mathbf{g} - \frac{\delta}{2}\mathbf{d}^T\mathsf{A}\mathbf{d} : \ t \in [0, 1]\} \\
&\leq \delta^2\lambda_{\min}\mathbf{d}^T\mathbf{g} + \frac{1}{2}\delta^2\lambda_{\min}^2\|\mathbf{d}\|^2 - \delta\mathbf{d}^T\mathbf{g} - \frac{\delta}{2}\mathbf{d}^T\mathsf{A}\mathbf{d} \\
&\leq \delta^2\lambda_{\min}\mathbf{d}^T\mathbf{g} + \frac{1}{2}\delta^2\lambda_{\min}\mathbf{d}^T\mathsf{A}\mathbf{d} - \delta\mathbf{d}^T\mathbf{g} - \frac{\delta}{2}\mathbf{d}^T\mathsf{A}\mathbf{d} \\
&= \delta(\delta\lambda_{\min} - 1)(\mathbf{d}^T\mathbf{g} + \frac{1}{2}\mathbf{d}^T\mathsf{A}\mathbf{d}) \\
&= \delta(\delta\lambda_{\min} - 1)\left(f(\mathbf{x} + \mathbf{d}) - f(\mathbf{x})\right) \\
&= \delta(1 - \delta\lambda_{\min})\left(f(\mathbf{x}) - f(\widehat{\mathbf{x}})\right).
\end{aligned}$$

□

Now we are ready to formulate and prove the main result of this section.

Proposition 6.3 *Let Ω be a product of ellipses and/or half-spaces, let $\widehat{\mathbf{x}}$ denote the unique solution of (6.1), $\mathbf{g} = \nabla f(\mathbf{x})$, $\mathbf{x} \in \Omega$, and let $\lambda_{\min}$ denote the smallest eigenvalue of A.*

If $\alpha \in (0, 2\|\mathsf{A}\|^{-1}]$, then

$$f\left(P_\Omega(\mathbf{x} - \alpha\mathbf{g})\right) - f(\widehat{\mathbf{x}}) \leq \eta\left(f(\mathbf{x}) - f(\widehat{\mathbf{x}})\right), \tag{6.28}$$

where

$$\eta = \eta(\alpha) = 1 - \widehat{\alpha}\lambda_{\min} \tag{6.29}$$

is the cost function reduction coefficient and $\widehat{\alpha} = \min\{\alpha, 2\|\mathsf{A}\|^{-1} - \alpha\}$.

Proof Let us first assume that $0 < \alpha\|\mathsf{A}\| \leq 1$ and let $\mathbf{x} \in \Omega$ be arbitrary but fixed, so that we can use Lemma 6.3 with $\delta = \alpha$ to get

$$F_\alpha\left(P_\Omega(\mathbf{x} - \alpha\mathbf{g})\right) - \alpha f(\widehat{\mathbf{x}}) \leq \alpha(1 - \alpha\lambda_{\min})\left(f(\mathbf{x}) - f(\widehat{\mathbf{x}})\right). \tag{6.30}$$

In combination with (6.24), this proves (6.28) for $0 < \alpha \leq \|\mathsf{A}\|^{-1}$.

To prove the statement for $\alpha \in (\|A\|^{-1}, 2\|A\|^{-1}]$, let us first assume that $\|A\| = 1$ and let $\alpha = 2 - \delta$, $\delta \in (0, 1)$. Then F_1 dominates f and

$$\delta F_1(\mathbf{y}) \leq \delta F_1(\mathbf{y}) + \frac{1-\delta}{2}\|\mathbf{y} - \mathbf{x}\|^2 = F_\delta(\mathbf{y}). \tag{6.31}$$

Thus, we can apply (6.23), Proposition 6.2, and the latter inequality to get

$$\begin{aligned} \delta f\big(P_\Omega\,(\mathbf{x} - \alpha\mathbf{g})\big) &\leq \delta F_1\big(P_\Omega\,(\mathbf{x} - \alpha\mathbf{g})\big) \leq \delta F_1\big(P_\Omega\,(\mathbf{x} - \delta\mathbf{g})\big) \\ &\leq F_\delta\big(P_\Omega\,(\mathbf{x} - \delta\mathbf{g})\big). \end{aligned}$$

Combining the latter inequalities with (6.30) for $\alpha = \delta$, we get

$$\delta f\big(P_\Omega\,(\mathbf{x} - \alpha\mathbf{g})\big) - \delta f(\widehat{\mathbf{x}}) \leq \delta(1 - \delta\lambda_{\min})\big(f(\mathbf{x}) - f(\widehat{\mathbf{x}})\big).$$

This proves the statement for $\alpha \in (\|A\|^{-1}, 2\|A\|^{-1})$ and $\|A\| = 1$. To finish the proof, apply the last inequality divided by δ to the function $\|A\|^{-1} f$ and recall that f and P_Ω are continuous. □

The estimate (6.28) gives the best value

$$\eta^{opt} = 1 - \kappa(\mathsf{A})^{-1}$$

for $\alpha = \|\mathsf{A}\|^{-1}$ with $\kappa(\mathsf{A}) = \|\mathsf{A}\|\|\mathsf{A}^{-1}\|$.

6.6 Comments and References

While the contraction properties of the Euclidean projection have been known for a long time (see, e.g., Bertsekas [3]), it seems that the first results concerning the decrease of the cost function along the projected-gradient path are due to Schöberl [7–9], who found the bound on the rate of convergence of the cost function in the energy norm for the gradient projection method with the fixed steplength $\alpha \in (0, \|\mathsf{A}\|^{-1}]$ in terms of bounds on the spectrum of the Hessian matrix A. Later Kučera [10] observed that the arguments provided by Schoberl are valid for any convex set.

A successful application of the projections in the energy norm with a longer steplength to the solution of contact problems by Farhat and Lesoinne in their FETI–DP code motivated the extension of the estimate for the bound constraints to $\alpha \in [0, 2\|\mathsf{A}\|^{-1}]$ by Dostál [5]. The latter proof used a simple property of interval that was generalized to the concept of subsymmetric set. Bouchala and Vodstrčil in [6] and [11] extended the estimates to the spheres and ellipses by proving that they are subsymmetric. Let us mention that it is known that there are convex sets that are not subsymmetric [11], but no example of a convex set for which the estimates do not hold is known.

The estimates of the decrease of quadratic cost function along the projected-gradient path are important for the development of effective monotonic algorithms that combine Euclidean projections and conjugate gradients – see Chaps. 7 and 8. Let us recall that the linear rate of convergence of the cost function for the gradient projection method was proved earlier even for more general problems by Luo and Tseng [4], but they did not make any attempt to specify the constants.

Very good experimental results were obtained by the so-called *fast gradient methods* that use a longer steplength $\alpha(\mathbf{x}^k) \in (\lambda_{\max}^{-1}, \lambda_{\min}^{-1})$ defined by various rules. Though these methods can guarantee neither the decrease of the cost function nor a faster rate of convergence, their potential should not be overlooked. The idea of the fast gradient algorithms originates in the pioneering paper by Barzilai and Borwein [12]. For the algorithms and convergence results concerning bound constrained problems, see, e.g., Birgin, Martínez, and Raydan [13, 14], Dai and Fletcher [15], and Grippo, Lampariello, and Lucidi [16]. For the experimental results obtained by the spectral gradient method with a fall-back, see Pospíšil and Dostál [17].

References

1. Saad, Y.: Iterative Methods for Large Linear Systems. SIAM, Philadelphia (2002)
2. Dostál, Z.: Optimal Quadratic Programming Algorithms, with Applications to Variational Inequalities, 1st edn. Springer, New York (2009)
3. Bertsekas, D.P.: Nonlinear Optimization. Athena Scientific, Belmont (1999)
4. Luo, Z.Q., Tseng, P.: Error bounds and convergence analysis of feasible descent methods: a general approach. Annal. Oper. Res. **46**, 157–178 (1993)
5. Dostál, Z.: On the decrease of a quadratic function along the projected-gradient path. ETNA **31**, 25–59 (2008)
6. Bouchala, J., Dostál, Z., Vodstrčil, P.: Separable spherical constraints and the decrease of a quadratic function in the gradient projection. JOTA **157**, 132–140 (2013)
7. Schöberl, J.: Solving the Signorini problem on the basis of domain decomposition techniques. Computing **60**(4), 323–344 (1998)
8. Dostál, Z., Schöberl, J.: Minimizing quadratic functions over non-negative cone with the rate of convergence and finite termination. Comput. Optim. Appl. **30**(1), 23–44 (2005)
9. Schöberl, J.: Efficient contact solvers based on domain decomposition techniques. Comput. Math. Appl. **42**, 1217–1228 (2001)
10. Kučera, R.: Convergence rate of an optimization algorithm for minimizing quadratic functions with separable convex constraints. SIAM J. Optim. **19**, 846–862 (2008)
11. Bouchala, J., Dostál, Z., Kozubek, T., Pospíšil, L., Vodstrčil, P.: On the solution of convex QPQC problems with elliptic and other separable constraints. Appl. Math. Comput. **247**(15), 848–864 (2014)
12. Barzilai, J., Borwein, J.M.: Two point step size gradient methods. IMA J. Numer. Anal. **8**, 141–148 (1988)
13. Birgin, E.G., Martínez, J.M., Raydan, M.: Nonmonotone spectral projected gradient methods on convex sets. SIAM J. Optim. **10**, 1196–1211 (2000)
14. Birgin, E.G., Raydan, M., Martínez, J.M.: Spectral projected gradient methods: review and perspectives. J. Stat. Softw. **60**, 3 (2014)
15. Dai, Y.H., Fletcher, R.: Projected Barzilai–Borwein methods for large-scale box-constrained quadratic programming. Numer. Math. **100**, 21–47 (2005)

16. Grippo, L., Lampariello, F., Lucidi, S.: A nonmonotone line search technique for Newton's method. SIAM J. Numer. Anal. **23**(4), 707–716 (1986)
17. Pospíšil, L., Dostál, Z.: The projected Barzilai–Borwein method with fall-back for strictly convex QCQP problems with separable constraints (to appear)

Chapter 7
MPGP for Separable QCQP

We are concerned with the special *QCQP problems* to find

$$\min_{\mathbf{x}\in\Omega_S} f(\mathbf{x}), \tag{7.1}$$

where $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathsf{A}\mathbf{x} - \mathbf{x}^T\mathbf{b}$, $\mathbf{b} \in \mathbb{R}^n$, $\mathsf{A} \in \mathbb{R}^{n\times n}$ is SPD,

$$\Omega_S = \Omega_1 \times \cdots \times \Omega_s, \quad \Omega_i = \{\mathbf{x}_i \in \mathbb{R}^{m_i} : h_i(\mathbf{x}_i) \le 0\}, \quad i \in \mathscr{S}, \quad \mathscr{S} = \{1, 2, \ldots, s\}, \tag{7.2}$$

and h_i are convex functions. We are especially interested in the problems where Ω_i denotes a half-interval, i.e.,

$$m_i = 1 \quad \text{and} \quad h_i(x_i) = \ell_i - x_i, \quad x_i, \ell_i \in \mathbb{R},$$

or an ellipse, i.e.,

$$m_i = 2, \quad h_i(\mathbf{x}_i) = (\mathbf{x}_i - \mathbf{y}_i)^T\mathsf{H}_i(\mathbf{x}_i - \mathbf{y}_i) - c_i, \quad \mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^2, \quad c_i > 0, \quad \mathsf{H}_i \; SPD.$$

To include the possibility that not all components of $\mathbf{x}$ are constrained, we admit $h_i(\mathbf{x}_i) = -1$. Problem (7.1) arises in the solution of transient contact problems or in the inner loop of TFETI-based algorithms for the solution of contact problems with possibly orthotropic friction.

Here we are interested in efficient algorithms for solving problems with large n and sparse and well-conditioned A. Such algorithms should be able to return an approximate solution at the cost proportional to the dimension n and to recognize an acceptable solution when it is found, as our goal here is also to solve the auxiliary problems generated by the algorithms for solving more general QCQP problems. Our choice is a variant of the active set strategy that we coined *MPGP* (Modified Proportioning with Gradient Projections). The algorithm uses conjugate gradients to solve auxiliary unconstrained problems with the precision controlled by the norm of violation of the Karush–Kuhn–Tucker conditions. The fixed step length gradient projections are used to change the active set.

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3_7

7.1 Projected Gradient, Reduced Gradient, and KKT Conditions

By Proposition 3.11, the solution to problem (7.1) always exists and is necessarily unique. As Ω_S satisfies ACQ, the unique solution $\widehat{\mathbf{x}}$ of (7.1) is fully determined by the KKT conditions (3.39), so that there is $\lambda \in \mathbb{R}^s$ such that

$$\widehat{\mathbf{g}}_i + \nabla h_i(\widehat{\mathbf{x}}_i)\lambda_i = \mathbf{o}, \quad h_i(\widehat{\mathbf{x}}_i)\lambda_i = 0, \quad \lambda_i \geq 0, h_i(\widehat{\mathbf{x}}_i) \leq 0, \ i = 1, \ldots, s, \tag{7.3}$$

where we use the notation $\widehat{\mathbf{g}} = \mathbf{g}(\widehat{\mathbf{x}})$.

To link the violation of the KKT conditions (7.3) to the solution error, we begin with some notations. Let $\mathscr{S}$ denote the set of all indices of the constraints so that

$$\mathscr{S} = \{1, 2, \ldots, s\}.$$

For any $\mathbf{x} \in \mathbb{R}^n$, we define the *active set* of $\mathbf{x}$ by

$$\mathscr{A}(\mathbf{x}) = \{i \in \mathscr{S} : h_i(\mathbf{x}_i) = 0\}.$$

Its complement

$$\mathscr{F}(\mathbf{x}) = \{i \in \mathscr{S} : h_i(\mathbf{x}_i) \neq 0\}$$

is called a *free set*.

For $\mathbf{x} \in \Omega_S$, we define the outer unit normal $\mathbf{n}$ by

$$\mathbf{n}_i = \mathbf{n}_i(\mathbf{x}) = \begin{cases} \|\nabla h_i(\mathbf{x}_i)\|^{-1} \nabla h_i(\mathbf{x}_i) & \text{for} \quad i \in \mathscr{A}(\mathbf{x}), \\ \mathbf{o} & \text{for} \quad i \in \mathscr{F}(\mathbf{x}). \end{cases}$$

The components of the gradient that violate the KKT conditions (7.3) in the free set and the active set are called a *free gradient* $\boldsymbol{\varphi}$ and a *chopped gradient* $\boldsymbol{\beta}$, respectively. They are defined by

$$\boldsymbol{\varphi}_i(\mathbf{x}) = \mathbf{g}_i(\mathbf{x}) \text{ for } i \in \mathscr{F}(\mathbf{x}), \quad \boldsymbol{\varphi}_i(\mathbf{x}) = \mathbf{o} \text{ for } i \in \mathscr{A}(\mathbf{x}) \tag{7.4}$$

$$\boldsymbol{\beta}_i(\mathbf{x}) = \mathbf{o} \ \text{ for } i \in \mathscr{F}(\mathbf{x}), \qquad \boldsymbol{\beta}_i(\mathbf{x}) = \mathbf{g}_i(\mathbf{x}) - (\mathbf{n}_i^T \mathbf{g}_i)^- \mathbf{n}_i \text{ for } i \in \mathscr{A}(\mathbf{x}), \tag{7.5}$$

where we use the notation $\mathbf{g}_i = \mathbf{g}_i(\mathbf{x})$ and

$$(\mathbf{n}_i^T \mathbf{g}_i)^- = \min\{\mathbf{n}_i^T \mathbf{g}_i, 0\}.$$

Thus the KKT conditions (7.3) are satisfied if and only if the *projected gradient*

$$\mathbf{g}^P(\mathbf{x}) = \boldsymbol{\varphi}(\mathbf{x}) + \boldsymbol{\beta}(\mathbf{x})$$

is equal to zero. Notice that $\mathbf{g}^P$ is not a continuous function of $\mathbf{x}$ in Ω_S. If $h_i(x_i)$ defines the bound constraint $x_i \geq \ell_i$, then $n_i = -1$ and the projected gradient is defined by

$$g_i^P = g_i^- \quad \text{for} \quad i \in \mathscr{A}(\mathbf{x}), \quad g_i^P = g_i \quad \text{for} \quad i \in \mathscr{F}(\mathbf{x}). \tag{7.6}$$

Since

$$\mathbf{g}_i^P(\mathbf{x}) = \mathbf{g}_i(\mathbf{x}), \quad i \in \mathscr{F}(\mathbf{x}),$$

and for any $i \in \mathscr{A}(\mathbf{x})$

$$\begin{aligned} \|\mathbf{g}_i^P\|^2 = \|\boldsymbol{\beta}_i\|^2 &= (\mathbf{g}_i - \{\mathbf{n}_i^T\mathbf{g}_i\}^-\mathbf{n}_i)^T(\mathbf{g}_i - \{\mathbf{n}_i^T\mathbf{g}_i\}^-\mathbf{n}_i) \\ &= \|\mathbf{g}_i\|^2 - \left(\{\mathbf{n}_i^T\mathbf{g}_i\}^-\right)^2 = \mathbf{g}_i^T\mathbf{g}_i^P, \end{aligned}$$

we have

$$\|\mathbf{g}^P\|^2 = \mathbf{g}^T\mathbf{g}^P \leq \|\mathbf{g}\|\,\|\mathbf{g}^P\| \tag{7.7}$$

and

$$\|\mathbf{g}^P\| \leq \|\mathbf{g}\|. \tag{7.8}$$

We need yet another simple property of the projected gradient.

Lemma 7.1 *Let* $\mathbf{x}, \mathbf{y} \in \Omega_S$ *and* $\mathbf{g} = \nabla f(\mathbf{x})$. *Then*

$$\mathbf{g}^T(\mathbf{y} - \mathbf{x}) \geq (\mathbf{g}^P)^T(\mathbf{y} - \mathbf{x}). \tag{7.9}$$

Proof First observe that

$$\mathbf{g}^T(\mathbf{y} - \mathbf{x}) = (\mathbf{g} - \mathbf{g}^P)^T(\mathbf{y} - \mathbf{x}) + (\mathbf{g}^P)^T(\mathbf{y} - \mathbf{x}).$$

Using the definition of projected gradient, we get

$$(\mathbf{g} - \mathbf{g}^P)^T(\mathbf{y} - \mathbf{x}) = \sum_{i \in \mathscr{S}} (\mathbf{g}_i - \mathbf{g}_i^P)^T(\mathbf{y}_i - \mathbf{x}_i) = \sum_{i \in \mathscr{A}(x)} (\mathbf{n}_i^T\mathbf{g}_i)^-\mathbf{n}_i^T(\mathbf{y}_i - \mathbf{x}_i).$$

To finish the proof, it is enough to observe that for $i \in \mathscr{A}(\mathbf{x})$

$$\mathbf{n}_i^T(\mathbf{y}_i - \mathbf{x}_i) \leq 0$$

due to the convexity of Ω_i. □

The following lemma can be considered as a quantitative refinement of the KKT conditions.

Lemma 7.2 *Let $\widehat{\mathbf{x}}$ be the solution of (7.1) and let $\mathbf{g}^P(\mathbf{x})$ denote the projected gradient at $\mathbf{x} \in \Omega_S$. Then*

$$\|\mathbf{x} - \widehat{\mathbf{x}}\|_{\mathsf{A}}^2 \le 2\big(f(\mathbf{x}) - f(\widehat{\mathbf{x}})\big) \le \|\mathbf{g}^P(\mathbf{x})\|_{\mathsf{A}^{-1}}^2 \le \lambda_{\min}(\mathsf{A})^{-1}\|\mathbf{g}^P(\mathbf{x})\|^2. \tag{7.10}$$

Proof Let $\widehat{\mathscr{A}}$, $\widehat{\mathscr{F}}$, and $\widehat{\mathbf{g}}$ denote the active set, free set, and the gradient at the solution, respectively. Observe that if $i \in \widehat{\mathscr{A}}$ and $\mathbf{x}_i \in \Omega_i$, then, using the convexity of h_i,

$$\big(\nabla h_i(\widehat{\mathbf{x}}_i)\big)^T(\mathbf{x}_i - \widehat{\mathbf{x}}_i) \le h_i(\mathbf{x}_i) - h_i(\widehat{\mathbf{x}}_i) = h_i(\mathbf{x}_i) \le 0.$$

It follows by the KKT conditions (7.3) and $\widehat{\mathbf{g}}_{\widehat{\mathscr{F}}} = \mathbf{o}_{\widehat{\mathscr{F}}}$ that

$$\widehat{\mathbf{g}}^T(\mathbf{x} - \widehat{\mathbf{x}}) = \sum_{i \in \widehat{\mathscr{A}}} \widehat{\mathbf{g}}_i^T(\mathbf{x}_i - \widehat{\mathbf{x}}_i) = \sum_{i \in \widehat{\mathscr{A}}} -\lambda_i \big(\nabla h_i(\widehat{\mathbf{x}}_i)\big)^T(\mathbf{x}_i - \widehat{\mathbf{x}}_i) \ge 0. \tag{7.11}$$

Thus, for any $\mathbf{x} \in \Omega_S$,

$$f(\mathbf{x}) - f(\widehat{\mathbf{x}}) = \widehat{\mathbf{g}}^T(\mathbf{x} - \widehat{\mathbf{x}}) + \frac{1}{2}(\mathbf{x} - \widehat{\mathbf{x}})^T\mathsf{A}(\mathbf{x} - \widehat{\mathbf{x}}) \ge \frac{1}{2}\|\mathbf{x} - \widehat{\mathbf{x}}\|_{\mathsf{A}}^2.$$

This proves the left inequality of (7.10).

To prove the right inequality, we can use Lemma 7.1 and simple manipulations to get for any $\mathbf{x} \in \Omega_S$

$$\begin{aligned} 0 &\ge 2\big(f(\widehat{\mathbf{x}}) - f(\mathbf{x})\big) = \|\widehat{\mathbf{x}} - \mathbf{x}\|_{\mathsf{A}}^2 + 2\mathbf{g}^T(\widehat{\mathbf{x}} - \mathbf{x}) \\ &\ge \|\widehat{\mathbf{x}} - \mathbf{x}\|_{\mathsf{A}}^2 + 2\left(\mathbf{g}^P\right)^T(\widehat{\mathbf{x}} - \mathbf{x}) \\ &\ge 2\min_{\mathbf{y} \in \mathbb{R}^n}\left(\frac{1}{2}\mathbf{y}^T\mathsf{A}\mathbf{y} + \left(\mathbf{g}^P\right)^T\mathbf{y}\right) = -(\mathbf{g}^P)^T\mathsf{A}^{-1}\mathbf{g}^P. \end{aligned}$$

The right inequality of (7.10) now follows easily. □

The projected gradient $\mathbf{g}^P$ is a natural error measure in the energy norm, but it can be very sensitive to the curvature of the boundary, as illustrated in Fig. 7.1.

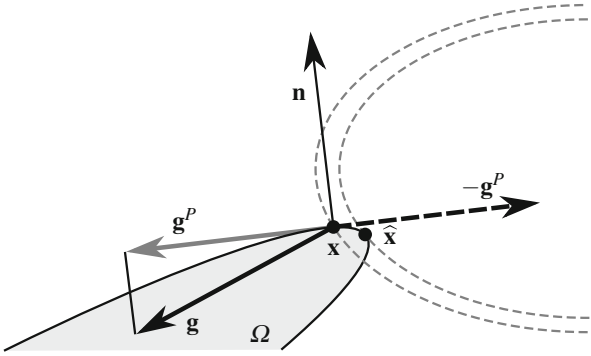

Fig. 7.1 Large projected gradient near the solution

If the curvature of the active constraints near the solution is strong as compared with the norm of the gradient, which is typical for the elliptic constraints near a vertex or for the spherical constraints with a small radius, then it is necessary to use more robust error measures, such as the *reduced gradient* $\widetilde{\mathbf{g}}_\alpha = \widetilde{\mathbf{g}}_\alpha(\mathbf{x})$, which is defined for any $\mathbf{x} \in \Omega_S$ and $\alpha > 0$ by

$$\widetilde{\mathbf{g}}_\alpha = \frac{1}{\alpha}\left(\mathbf{x} - P_{\Omega_S}(\mathbf{x} - \alpha\mathbf{g})\right). \tag{7.12}$$

It is easy to check that the reduced gradient $\widetilde{\mathbf{g}}_\alpha$ is a continuous function of $\mathbf{x}$ and α. The following lemma is an alternative to Lemma 7.2.

Lemma 7.3 *Let $\widehat{\mathbf{x}}$ be the solution of (7.1) and let $\widetilde{\mathbf{g}}_\alpha = \widetilde{\mathbf{g}}_\alpha(\mathbf{x})$ denote the reduced gradient at $\mathbf{x} \in \Omega_S$.*

Then

$$\|\mathbf{x} - \widehat{\mathbf{x}}\| \le \nu(\alpha)\|\widetilde{\mathbf{g}}_\alpha\|, \tag{7.13}$$

where

$$\nu(\alpha) = \begin{cases} (\lambda_{\min}(\mathsf{A}))^{-1} & \text{for } \ 0 < \alpha \le 2(\lambda_{\min}(\mathsf{A}) + \|\mathsf{A}\|)^{-1}, \\ \alpha\,(2 - \alpha\|\mathsf{A}\|)^{-1} & \text{for } \ 2(\lambda_{\min}(\mathsf{A}) + \|\mathsf{A}\|)^{-1} \le \alpha < 2\|\mathsf{A}\|^{-1}. \end{cases} \tag{7.14}$$

Proof Using the properties of norm and Corollary 3.1, we get that for any $\alpha > 0$

$$\begin{aligned} \|\mathbf{x} - \widehat{\mathbf{x}}\| &\le \|P_{\Omega_S}(\mathbf{x} - \alpha\mathbf{g}) - \mathbf{x}\| + \|P_{\Omega_S}(\mathbf{x} - \alpha\mathbf{g}) - \widehat{\mathbf{x}}\| \\ &\le \alpha\|\widetilde{\mathbf{g}}_\alpha\| + \|P_{\Omega_S}(\mathbf{x} - \alpha\mathbf{g}) - P_{\Omega_S}\left(\widehat{\mathbf{x}} - \alpha\mathbf{g}(\widehat{\mathbf{x}})\right)\| \\ &\le \alpha\|\widetilde{\mathbf{g}}_\alpha\| + \max\{|1 - \alpha\lambda_{\min}(\mathsf{A})|, |1 - \alpha\lambda_{\max}(\mathsf{A})|\}\|\mathbf{x} - \widehat{\mathbf{x}}\|. \end{aligned}$$

After simple manipulations, we get (7.13). □

We shall also need the inequalities formulated in the following lemma.

Lemma 7.4 *Let* $\mathbf{x} \in \Omega_S$, $\alpha > 0$, *and let us denote* $\mathbf{g} = \mathbf{g}(\mathbf{x}) = \mathsf{A}\mathbf{x} - \mathbf{b}$, $\widetilde{\mathbf{g}}_\alpha = \widetilde{\mathbf{g}}_\alpha(\mathbf{x})$, *and* $\mathbf{g}^P = \mathbf{g}^P(\mathbf{x})$. *Then*

$$\|\widetilde{\mathbf{g}}_\alpha\|^2 \le \mathbf{g}^T\widetilde{\mathbf{g}}_\alpha \le \mathbf{g}^T\mathbf{g}^P = \|\mathbf{g}^P\|^2 \le \|\mathbf{g}\|^2. \tag{7.15}$$

Proof Using that for all $\mathbf{x} \in \Omega_S$ and $\mathbf{y} \in \mathbb{R}^n$

$$\big(\mathbf{x} - P_{\Omega_S}(\mathbf{y})\big)^T\big(\mathbf{y} - P_{\Omega_S}(\mathbf{y})\big) \le 0$$

by Proposition 3.5, we get for any $\alpha > 0$ and $\mathbf{y} = \mathbf{x} - \alpha\mathbf{g}$

$$\alpha\widetilde{\mathbf{g}}_\alpha^T(-\alpha\mathbf{g} + \alpha\widetilde{\mathbf{g}}_\alpha) = \big(\mathbf{x} - (\mathbf{x} - \alpha\widetilde{\mathbf{g}}_\alpha)\big)^T\big(\mathbf{x} - \alpha\mathbf{g} - (\mathbf{x} - \alpha\widetilde{\mathbf{g}}_\alpha)\big) \le 0.$$

After simple manipulations, we get the left inequality of (7.15).

To prove the rest, notice that $\mathbf{x} - \mathbf{g}^P$ is the projection of $\mathbf{x} - \mathbf{g}$ to the set

$$\hat{\Omega}(\mathbf{x}) = \hat{\Omega} = \hat{\Omega}_1 \times \cdots \times \hat{\Omega}_s,$$

where

$$\begin{aligned}
&\hat{\Omega}_i = \{\mathbf{y}_i \in \mathbb{R}^{\ell_i} : \hat{h}_i(\mathbf{y}_i) \le 0\},\\
&\hat{h}_i(\mathbf{y}_i) = (\mathbf{y}_i - \mathbf{x}_i)^T\nabla h_i(\mathbf{x}_i) && \text{for } i \in \mathscr{A}(\mathbf{x}),\\
&\hat{h}_i(\mathbf{y}_i) = -1 && \text{for } i \in \mathscr{F}(\mathbf{x}).
\end{aligned}$$

Since $\mathbf{x} - \alpha\widetilde{\mathbf{g}}_\alpha$ is the projection of $\mathbf{x} - \alpha\mathbf{g}$ to Ω_S, $\mathbf{x} - \alpha\mathbf{g}^P$ is the projection of $\mathbf{x} - \alpha\mathbf{g}$ to $\hat{\Omega}_S$, and $\Omega_S \subseteq \hat{\Omega}$, we have $\mathbf{x} - \alpha\widetilde{\mathbf{g}}_\alpha \in \Omega_S$ and

$$\alpha^2\|\mathbf{g} - \mathbf{g}^P\|^2 = \|(\mathbf{x} - \alpha\mathbf{g}) - (\mathbf{x} - \alpha\mathbf{g}^P)\|^2 \le \|(\mathbf{x} - \alpha\mathbf{g}) - (\mathbf{x} - \alpha\widetilde{\mathbf{g}}_\alpha)\|^2 = \alpha^2\|\mathbf{g} - \widetilde{\mathbf{g}}\|^2.$$

It follows that

$$-\mathbf{g}^T\mathbf{g}^P = -2\mathbf{g}^T\mathbf{g}^P + \|\mathbf{g}^P\|^2 \le -2\mathbf{g}^T\widetilde{\mathbf{g}} + \|\widetilde{\mathbf{g}}\|^2 \le -\mathbf{g}^T\widetilde{\mathbf{g}}.$$

The rest follows by (7.7) and (7.8). □

7.2 Reduced Projected Gradient

Though the reduced gradient provides a nice and robust estimate of the solution error, closer analysis reveals that it provides a little information about the local behavior of f in Ω_S. For example, the inequality

$$\mathbf{g}^T(\mathbf{y} - \mathbf{x}) \ge (\mathbf{g}^P)^T(\mathbf{y} - \mathbf{x}),$$

which is essential in the development of algorithms in Sect. 9.10, does not hold when we replace the projected gradient $\mathbf{g}^P$ by the reduced gradient $\widetilde{\mathbf{g}}_\alpha$. Moreover, there is no constant C such that $C\|\mathbf{g}^P(\mathbf{x})\| \leq \|\widetilde{\mathbf{g}}_\alpha(\mathbf{x})\|$ for $\mathbf{x} \in \Omega_S$. The reason is that the free components of $\widetilde{\mathbf{g}}_\alpha(\mathbf{x})$ are different from the corresponding components of $\mathbf{g}(\mathbf{x})$, which defines the best linear approximation of the change of $f(\mathbf{x})$ near $\mathbf{x}$ in terms of the variation of free variables.

The remedy is the *reduced projected gradient*

$$\widetilde{\mathbf{g}}^P_\alpha(\mathbf{x}) = \boldsymbol{\varphi}(\mathbf{x}) + \widetilde{\boldsymbol{\beta}}^\alpha(\mathbf{x}),$$

which is defined for each $\mathbf{x} \in \Omega$ by the free gradient $\boldsymbol{\varphi}(\mathbf{x})$ and the *reduced chopped gradient* $\widetilde{\boldsymbol{\beta}}^\alpha(\mathbf{x})$ with the components

$$\boldsymbol{\varphi}_i(\mathbf{x}) = \mathbf{g}_i(\mathbf{x}) \text{ for } i \in \mathscr{F}(\mathbf{x}), \quad \boldsymbol{\varphi}_i(\mathbf{x}) = \mathbf{o} \text{ for } i \in \mathscr{A}(\mathbf{x}), \tag{7.16}$$

$$\widetilde{\boldsymbol{\beta}}^\alpha_i(\mathbf{x}) = \begin{cases} \mathbf{o} \text{ for } i \in \mathscr{F}(\mathbf{x}), \\ \mathbf{g}_i(\mathbf{x}) \text{ for } i \in \mathscr{A}(\mathbf{x}) \text{ and } \mathbf{n}_i^T\mathbf{g}_i > 0, \\ \frac{1}{\alpha}\big(\mathbf{x}_i - P_{\Omega_i}(\mathbf{x}_i - \alpha\mathbf{g}_i)\big) \text{ for } i \in \mathscr{A}(\mathbf{x}) \text{ and } \mathbf{n}_i^T\mathbf{g}_i \leq 0. \end{cases} \tag{7.17}$$

Observe that $\widetilde{\mathbf{g}}^P_\alpha$ is not a continuous function of $\mathbf{x}$ and the KKT conditions (7.3) are satisfied if and only if the reduced projected gradient is equal to zero.

Comparing the definitions of $\widetilde{\mathbf{g}}_\alpha$ and $\widetilde{\mathbf{g}}^P_\alpha$ and using (7.15), we get

$$\|\widetilde{\mathbf{g}}_\alpha\| \leq \|\widetilde{\mathbf{g}}^P_\alpha\| \leq \|\mathbf{g}^P\|. \tag{7.18}$$

We can also formulate the following lemma.

Lemma 7.5 *Let $\widehat{\mathbf{x}}$ be the solution of (7.1) and let $\widetilde{\mathbf{g}}^P_\alpha = \widetilde{\mathbf{g}}^P_\alpha(\mathbf{x})$ denote the reduced projected gradient at $\mathbf{x} \in \Omega_S$. Then*

$$\|\mathbf{x} - \widehat{\mathbf{x}}\| \leq \nu(\alpha)\|\widetilde{\mathbf{g}}^P_\alpha\| \leq \nu(\alpha)\|\mathbf{g}^P\|, \tag{7.19}$$

where $\nu(\alpha)$ is defined in Lemma 7.3.

Proof The inequalities are easy corollaries of Lemma 7.3 and (7.18).

Lemma 7.6 *Let the set Ω_S of problem (7.1) be defined by the convex functions h_i with continuous second derivatives and let $\widehat{\mathbf{x}}$ denote the solution of (7.1). Then there are constants $\varepsilon > 0$ and $C > 0$ such that if $\mathbf{x} \in \Omega$, $\|\mathbf{x} - \widehat{\mathbf{x}}\| \leq \varepsilon$, and $i \in \mathscr{S}$, then*

$$\|\mathbf{g}^P_i(\mathbf{x})\| \leq C\|\mathbf{x} - \widehat{\mathbf{x}}\|. \tag{7.20}$$

Proof It is enough to prove (7.20) for C dependent on i. Let us first consider the components associated with the free or weakly binding set of the solution

$$\mathscr{Z} = \{i \in \mathscr{S} : \mathbf{g}_i(\widehat{\mathbf{x}}) = \mathbf{o}\}.$$

Then

$$\|\mathbf{g}_i(\mathbf{x})\| = \|\mathbf{g}_i(\mathbf{x}) - \mathbf{g}_i(\widehat{\mathbf{x}})\| = \|\mathsf{A}_{i*}(\mathbf{x} - \widehat{\mathbf{x}})\| \le \|\mathsf{A}\|\|\mathbf{x} - \widehat{\mathbf{x}}\|,$$

where A_{i*} denotes the block row of A which corresponds to $\mathbf{g}_i$, $i \in \mathscr{Z}$. Observing that $\|\mathbf{g}_i^P(\mathbf{x})\| \le \|\mathbf{g}_i(\mathbf{x})\|$ by (7.7), we get

$$\|\mathbf{g}_i^P(\mathbf{x})\| \le C_i\|\mathbf{x} - \widehat{\mathbf{x}}\|$$

with $C = \|\mathsf{A}\|$ and $\varepsilon = \infty$.

Let us now consider the components of $\mathbf{g}(\mathbf{x})$ with indices in

$$\mathscr{B} = \{i \in \mathscr{A}(\mathbf{x}) : \mathbf{g}_i^T(\widehat{\mathbf{x}})\mathbf{n}_i(\widehat{\mathbf{x}}) < 0\},$$

so there is $\varepsilon_0 > 0$ such that $\mathbf{g}_i(\mathbf{x}) \ne \mathbf{o}$ for $\|\mathbf{x} - \widehat{\mathbf{x}}\| \le \varepsilon_0$, and notice that

$$\mathbf{g}_i^T(\widehat{\mathbf{x}})\nabla h_i(\widehat{\mathbf{x}}_i) = -\|\mathbf{g}_i^T(\widehat{\mathbf{x}})\|\|\nabla h_i(\widehat{\mathbf{x}}_i)\| < 0.$$

It follows that there is $\varepsilon \in (0, \varepsilon_0)$ such that if $\mathbf{x} \in \mathbb{R}^n$ satisfies $\|\mathbf{x} - \widehat{\mathbf{x}}\| \le \varepsilon$, then

$$\mathbf{g}_i^T(\mathbf{x})\nabla h_i(\mathbf{x}_i) < 0, \quad i \in \mathscr{B}.$$

Moreover, the mapping $\psi_i : \mathbb{R}^n \to \mathbb{R}^{\ell_i}$ defined for $\|\mathbf{x} - \widehat{\mathbf{x}}\| \le \varepsilon$ by

$$\psi_i(\mathbf{x}) = \mathbf{g}_i(\mathbf{x}) - \frac{\mathbf{g}_i^T(\mathbf{x})\nabla h_i(\mathbf{x}_i)}{\|\nabla h_i(\mathbf{x}_i)\|^2}\nabla h_i(\mathbf{x}_i)$$

is differentiable at $\mathbf{x}$ and for $i \in \mathscr{B}$

$$\psi_i(\mathbf{x}) = \mathbf{g}_i(\mathbf{x}) - \frac{(\mathbf{g}_i^T(\mathbf{x})\nabla h_i(\mathbf{x}_i))^-}{\|\nabla h_i(\mathbf{x}_i)\|^2}\nabla h_i(\mathbf{x}_i) = \mathbf{g}_i^P(\mathbf{x}).$$

It follows that we can use the classical results of calculus to get that for each component $i \in \mathscr{B}$, there is $C > 0$ such $\|\mathbf{x} - \widehat{\mathbf{x}}\| \le \varepsilon$ implies

$$\|\psi_i(\mathbf{x}) - \psi_i(\widehat{\mathbf{x}})\| \le C\|\mathbf{x} - \widehat{\mathbf{x}}\|.$$

To finish the proof, recall that $\psi(\widehat{\mathbf{x}}) = \mathbf{o}$ and that for $i \in \mathscr{B}$ and $\|\mathbf{x} - \widehat{\mathbf{x}}\| \le \varepsilon$

$$\mathbf{g}_i^P(\mathbf{x}) = \psi_i(\mathbf{x}). \qquad \square$$

Now we are ready to prove the main result of this section.

Theorem 7.1 *Let the feasible set $\Omega_S = \Omega_1 \times \cdots \times \Omega_s$ of (7.1) be defined by the convex functions $h_i : \mathbb{R}^{\ell_i} \to \mathbb{R}$, each with a continuous second derivative. Then there is $C > 0$ and $\varepsilon > 0$ such that for each $\mathbf{x} \in \Omega_S$ and $\alpha \in (0, 2\|\mathsf{A}\|^{-1})$*

$$\|\widetilde{\mathbf{g}}_\alpha^P(\mathbf{x})\| \le \|\mathbf{g}^P(\mathbf{x})\| \le C\|\widetilde{\mathbf{g}}_\alpha^P(\mathbf{x})\|. \tag{7.21}$$

Proof Let $\alpha \in (0, 2\|\mathsf{A}\|^{-1})$ and $\mathbf{x} \in \Omega_S$ be fixed, so that we can simplify the notation by

$$\mathbf{g} = \mathbf{g}(\mathbf{x}), \quad \widetilde{\mathbf{g}}^P = \widetilde{\mathbf{g}}_\alpha^P(\mathbf{x}), \quad \text{and} \quad \mathbf{g}^P = \mathbf{g}^P(\mathbf{x}).$$

The left inequality of (7.21) has already been established—see (7.18). We shall prove the right inequality of (7.21) componentwise, observing that by the definitions

$$\|\mathbf{g}_i^P\| = \|\widetilde{\mathbf{g}}_i^P\| \quad \text{for} \quad i \in \mathscr{F}(\mathbf{x}).$$

If $i \in \mathscr{A}(\mathbf{x})$ and $\varepsilon > 0$, $C > 0$ are those of Lemma 7.6, then $\|\mathbf{x} - \widehat{\mathbf{x}}\| \le \varepsilon$ and by Lemmas 7.5 and 7.6

$$\|\mathbf{g}_i^P\| \le C\|\mathbf{x} - \widehat{\mathbf{x}}\| \le \nu(\alpha)C\|\widetilde{\mathbf{g}}^P\|, \tag{7.22}$$

where $\nu(\alpha)$ is defined in Lemma 7.3. This proves the right inequality of (7.21) for $\mathbf{x}$ near the solution.

To prove (7.21) for $i \in \mathscr{A}(\mathbf{x})$ and $\|\mathbf{x} - \widehat{\mathbf{x}}\| \ge \varepsilon$, observe that

$$\|\mathbf{g}_i^P\| \le \|\mathbf{g}_i\| = \|\left(\mathbf{g}_i - \mathbf{g}_i(\widehat{\mathbf{x}})\right) + \mathbf{g}_i(\widehat{\mathbf{x}})\| \le \|\mathsf{A}\|\|\mathbf{x} - \widehat{\mathbf{x}}\| + C_1. \tag{7.23}$$

Moreover, by Lemma 7.5

$$\|\mathbf{x} - \widehat{\mathbf{x}}\| \le \nu(\alpha)\|\widetilde{\mathbf{g}}^P\|. \tag{7.24}$$

Thus for $\|\mathbf{x} - \widehat{\mathbf{x}}\| \ge \varepsilon$

$$\begin{aligned}\|\mathbf{g}_i^P\| &\le \|\mathsf{A}\|\|\mathbf{x} - \widehat{\mathbf{x}}\| + (C_1/\varepsilon)\varepsilon \le \left(\|\mathsf{A}\| + (C_1/\varepsilon)\right)\|\mathbf{x} - \widehat{\mathbf{x}}\| \\ &\le \left(\|\mathsf{A}\| + (C_1/\varepsilon)\right)\nu(\alpha)\|\widetilde{\mathbf{g}}^P\|. \end{aligned} \tag{7.25}$$

The rest follows by the finite dimension argument. □

7.3 MPGP Scheme

The algorithm that we propose here exploits a user-defined constant $\Gamma > 0$, a test which is used to decide when to change the face, and two types of steps.

The *conjugate gradient step* is defined by

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_{cg}\mathbf{p}^{k+1}, \quad \alpha_{cg} = \mathbf{b}^T\mathbf{p}^{k+1} / \left(\mathbf{p}^{k+1}\right)^T \mathsf{A}\mathbf{p}^{k+1}, \tag{7.26}$$

where $\mathbf{p}^{k+1}$ is the conjugate gradient direction (see Sect. 5.2) constructed recurrently. The recurrence starts (or restarts) with $\mathbf{p}^{k+1} = \boldsymbol{\varphi}(\mathbf{x}^k)$ whenever $\mathbf{x}^k$ is generated by the gradient projection step. If $\mathbf{x}^k$ is generated by the conjugate gradient step, then $\mathbf{p}^{k+1}$ is given by

$$\mathbf{p}^{k+1} = \boldsymbol{\varphi}(\mathbf{x}^k) - \gamma \mathbf{p}^k, \quad \gamma = \frac{\boldsymbol{\varphi}(\mathbf{x}^k)^T \mathsf{A}\mathbf{p}^k}{(\mathbf{p}^k)^T \mathsf{A}\mathbf{p}^k}. \tag{7.27}$$

The coefficient α_{cg} is chosen so that

$$\begin{aligned} f(\mathbf{x}^{k+1}) &= \min\{f\left(\mathbf{x}^k - \alpha \mathbf{p}^{k+1}\right) : \alpha \in \mathbb{R}\} \\ &= \min\{f(\mathbf{x}) : \mathbf{x} \in \mathbf{x}^r + \operatorname{Span}\left\{\mathbf{p}^{r+1}, \ldots, \mathbf{p}^{k+1}\right\}. \end{aligned}$$

It can be checked directly that

$$f(\mathbf{x}^{k+1}) \le f\left(\mathbf{x}^k - \alpha_{cg} \boldsymbol{\varphi}(\mathbf{x}^k)\right) = f(\mathbf{x}^k) - \frac{1}{2} \frac{\|\boldsymbol{\varphi}(\mathbf{x}^k)\|^4}{\boldsymbol{\varphi}(\mathbf{x}^k)^T \mathsf{A} \boldsymbol{\varphi}(\mathbf{x}^k)}. \tag{7.28}$$

The conjugate gradient steps are used to speed up the minimization in the face

$$\mathscr{W}_{\mathscr{J}} = \{\mathbf{x} : h_i(\mathbf{x}_i) = 0, \ \ i \in \mathscr{J}\}, \quad \mathscr{J} = \mathscr{A}(\mathbf{x}^k).$$

The *gradient projection step* is defined by the gradient projection

$$\mathbf{x}^{k+1} = P_{\Omega_S}\left(\mathbf{x}^k - \alpha \mathbf{g}(\mathbf{x}^k)\right) = \mathbf{x}^k - \alpha \widetilde{\mathbf{g}}_\alpha \tag{7.29}$$

with a fixed step length $\alpha > 0$. This step can both add and remove the indices from the current working set.

If for a given $\Gamma > 0$ the inequality

$$\|\boldsymbol{\beta}(\mathbf{x}^k)\| \le \Gamma \|\boldsymbol{\varphi}(\mathbf{x}^k)\| \tag{7.30}$$

holds, then we call the iterate $\mathbf{x}^k$ *proportional*. Test (7.30) is used to decide if the algorithm should continue exploration of the current working set by the conjugate gradients or change the working set by the gradient projection step. If $\Gamma = 1$, then the iterate is proportional if the free gradient dominates the violation of the KKT conditions.

Alternatively, test (7.30) can be written in the form

$$2\delta \|\mathbf{g}^P(\mathbf{x})\|^2 \le \|\boldsymbol{\varphi}(\mathbf{x}^k)\|^2 \tag{7.31}$$

with

$$\delta = \frac{1}{2\Gamma^2 + 2}, \quad \delta \in (0, 1/2),$$

which is more convenient for the analysis of algorithms. It is easy to check that the tests (7.30) and (7.31) are equivalent.

Now we are ready to describe the basic algorithm in the form which is convenient for the analysis. More details about the implementation and the choice of parameters can be found in Sect. 7.6.2.

Algorithm 7.1 Modified proportioning with gradient projections (MPGP schema).

Given an SPD matrix $\mathsf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, $\Gamma > 0$.
Choose $\mathbf{x}^0 \in \Omega_S$ *and* $\alpha \in (0, 2\|\mathsf{A}\|^{-1})$.
For $k = 0, 1, \ldots,$ *choose* $\mathbf{x}^{k+1}$ *by the following rules:*

(i) If $\mathbf{g}^P(\mathbf{x}^k) = \mathbf{o}$, *set* $\mathbf{x}^{k+1} = \mathbf{x}^k$.

(ii) If $\mathbf{x}^k$ *is proportional and* $\mathbf{g}^P(\mathbf{x}^k) \neq \mathbf{o}$, *try to generate* $\mathbf{x}^{k+1}$ *by the conjugate gradient step. If* $\mathbf{x}^{k+1} \in \Omega$, *then accept it, else generate* $\mathbf{x}^{k+1}$ *by the gradient projection step.*

(iii) If $\mathbf{x}^k$ *is not proportional, define* $\mathbf{x}^{k+1}$ *by the gradient projection step.*

Remark 7.1 The gradient projection step in (ii) can be replaced by the expansion step

$$\mathbf{x}^{k+1} = P_\Omega\left(\mathbf{x}^k - \alpha\boldsymbol{\varphi}(\mathbf{x}^k)\right).$$

7.4 Rate of Convergence

In this section, we give the bounds on the difference between the value of the cost function at the solution and the current iterate. Let us first examine the effect of the CG step.

Lemma 7.7 *Let* Ω_S *denote a closed convex set, let* $\widehat{\mathbf{x}}$ *denote a unique solution of (7.1), let* $\lambda_{\min}$ *denote the smallest eigenvalue of* A, *let* $\{\mathbf{x}^k\}$ *denote the iterates generated by Algorithm 7.1 with* $\delta \in (0, 1/2)$, *and let* $\mathbf{x}^{k+1}$ *be generated by the CG step. Then*

$$f\left(\mathbf{x}^{k+1}\right) - f(\widehat{\mathbf{x}}) \leq \eta\left(\delta\|\mathsf{A}\|^{-1}\right)\left(f(\mathbf{x}^k) - f(\widehat{\mathbf{x}})\right), \tag{7.32}$$

where

$$\eta(\xi) = 1 - \xi\lambda_{\min}.$$

Proof Let $\mathbf{x}^{k+1}$ be generated by the CG step, so that $\mathbf{x}^k$ is proportional (7.31), and denote $\alpha = \delta\|\mathsf{A}\|^{-1}$. Using (7.28), (7.31), (7.15), and simple manipulations, we get

$$
\begin{aligned}
f(\mathbf{x}^{k+1}) &= f\left(\mathbf{x}^k - \alpha_{cg}\boldsymbol{\varphi}(\mathbf{x}^k)\right) = f(\mathbf{x}^k) - \frac{1}{2}\frac{\|\boldsymbol{\varphi}(\mathbf{x}^k)\|^4}{\boldsymbol{\varphi}(\mathbf{x}^k)^T\mathsf{A}\boldsymbol{\varphi}(\mathbf{x}^k)} \\
&\le f(\mathbf{x}^k) - \frac{1}{2}\|\mathsf{A}\|^{-1}\|\boldsymbol{\varphi}(\mathbf{x}^k)\|^2 \le f(\mathbf{x}^k) - \alpha\mathbf{g}^T(\mathbf{x}^k)\mathbf{g}^P(\mathbf{x}^k) \\
&\le f(\mathbf{x}^k) - \alpha\widetilde{\mathbf{g}}_\alpha^T(\mathbf{x}^k)\mathbf{g}(\mathbf{x}^k) \\
&\le f(\mathbf{x}^k) - \alpha\widetilde{\mathbf{g}}_\alpha^T(\mathbf{x}^k)\mathbf{g}(\mathbf{x}^k) + \frac{\alpha^2}{2}\widetilde{\mathbf{g}}_\alpha^T(\mathbf{x}^k)\mathsf{A}\widetilde{\mathbf{g}}_\alpha(\mathbf{x}^k) \\
&= f\Big(P_\Omega\left(\mathbf{x}^k - \alpha\mathbf{g}(\mathbf{x}^k)\right)\Big).
\end{aligned}
$$

After subtracting $f(\widehat{\mathbf{x}})$ from the first and the last expression, we get

$$
f(\mathbf{x}^{k+1}) - f(\widehat{\mathbf{x}}) \le f\Big(P_\Omega\left(\mathbf{x}^k - \delta\|\mathsf{A}\|^{-1}\mathbf{g}(\mathbf{x}^k)\right)\Big) - f(\widehat{\mathbf{x}}). \tag{7.33}
$$

The rest follows by (6.28). □

The main result reads as follows.

Theorem 7.2 *Let Ω_S be a closed convex set, let $\widehat{\mathbf{x}}$ denote a unique solution of (7.1), let $\lambda_{\min}$ denote the smallest eigenvalue of A, and let $\{\mathbf{x}^k\}$ be generated by Algorithm 7.1 with $\mathbf{x}^0 \in \Omega$, $\alpha \in (0, 2\|\mathsf{A}\|^{-1})$, and $\delta \in (0, 1/2)$.*

Then for any $k \ge 0$

$$
f\left(\mathbf{x}^{k+1}\right) - f(\widehat{\mathbf{x}}) \le \eta\big(f(\mathbf{x}^k) - f(\widehat{\mathbf{x}})\big), \tag{7.34}
$$

where

$$
\eta = \eta(\delta, \alpha) = 1 - \delta\widehat{\alpha}\lambda_{\min} \le 1 - \delta\kappa(\mathsf{A})^{-1}, \quad \widehat{\alpha} = \min\{\alpha, 2\|\mathsf{A}\|^{-1} - \alpha\}. \tag{7.35}
$$

Proof If $\mathbf{x}^{k+1}$ is generated by the gradient projection step, then the (7.34) is satisfied by Proposition 6.3. If $\mathbf{x}^{k+1}$ is generated by the CG step, then (7.34) is satisfied by Lemma 7.7. □

7.5 Bound on Norm of Projected Gradient

To use the MPGP algorithm in the inner loops of other algorithms, we must be able to *recognize* when we are near the solution. However, there is a catch—though the latter can be tested by a norm of the projected gradient by Lemma 7.2, Theorem 7.2 does not guarantee that such test is positive near the solution. The projected gradient is not continuous and can be large near the solution as shown in Fig. 7.2.

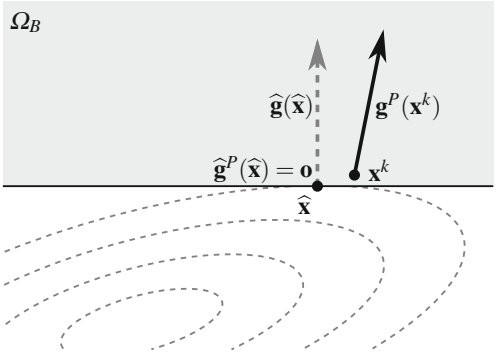

Fig. 7.2 Large projected gradient near the solution

Here we show that the situation is different for the subsequent iterates of MPGP. To see why, let us assume that $\{\mathbf{x}^k\}$ is generated by MPGP for the solution of (7.1) and let $k \geq 1$ be arbitrary but fixed. The main tool in our analysis is the linearized problem associated with $\mathbf{x}^k$ that reads

$$\min\ f(\mathbf{x}) \quad \text{subject to} \quad \mathbf{x} \in \hat{\Omega}^k, \quad \hat{\Omega}^k = \hat{\Omega}^k_1 \times \cdots \times \hat{\Omega}^k_s, \tag{7.36}$$

where

$$\begin{aligned} &\hat{\Omega}^k_i = \{\mathbf{x}_i \in \mathbb{R}^{\ell_i} : \hat{h}_i(\mathbf{x}_i) \leq 0\} && \text{for } i \in \mathscr{S}, \\ &\hat{h}_i(\mathbf{x}_i) = (\mathbf{x}_i - \mathbf{x}^k_i)^T \nabla h_i(\mathbf{x}^k_i) && \text{for } i \in \mathscr{A}(\mathbf{x}^k), \\ &\hat{h}_i(\mathbf{x}_i) = -1 && \text{for } i \in \mathscr{F}(\mathbf{x}^k). \end{aligned}$$

Comparing (7.36) with our original problem (7.1), we can see that the original constraints on $\mathbf{x}_i$ are omitted in (7.36) for $i \in \mathscr{F}(\mathbf{x}^k)$ and replaced by their linearized versions for $i \in \mathscr{A}(\mathbf{x}^k)$. Since h_i are convex by the assumptions, we get easily

$$\Omega_S \subseteq \hat{\Omega} \text{ and } \mathbf{n}_i = \hat{\mathbf{n}}_i,\ i \in \mathscr{A}(\mathbf{x}^k). \tag{7.37}$$

Problem (7.36) is defined so that the iterate $\mathbf{x}^k$, obtained from $\mathbf{x}^{k-1}$ by MPGP for the solution of problem (7.1), can also be considered as an iterate for the solution of (7.36). We use the hat to distinguish the concepts related to problem (7.36) from those related to the original problem (7.1). For example, $\hat{\mathscr{A}}^k(\mathbf{x})$ denotes the active set of $\mathbf{x} \in \mathbb{R}^n$ with respect to $\hat{\Omega}^k$. For typographical reasons, we denote the reduced gradient for (7.36) by $\hat{\mathbf{g}}_\alpha$. The following relations are important in what follows.

Lemma 7.8 *Let $\mathbf{x}^k$ denote an iterate generated by the MPGP algorithm under the assumptions of Theorem 7.2, let problem (7.36) be associated with $\mathbf{x}^k$, and let $\hat{\mathbf{g}}^P(\mathbf{x}^k)$ and $\hat{\mathbf{g}}_\alpha(\mathbf{x}^k)$ denote the projected gradient and the reduced gradient associated with problem (7.36), respectively. Then*

$$\mathbf{g}^P(\mathbf{x}^k) = \hat{\mathbf{g}}^P(\mathbf{x}^k) = \hat{\mathbf{g}}_\alpha(\mathbf{x}^k). \tag{7.38}$$

Proof Let $i \in \mathscr{A}(\mathbf{x}^k)$, $\mathbf{n} = \mathbf{n}(\mathbf{x}^k)$, $\mathbf{g} = \mathbf{g}(\mathbf{x}^k)$, $\alpha > 0$, $\hat{\mathbf{g}} = \hat{\mathbf{g}}_\alpha(\mathbf{x}^k)$, and $\mathbf{n}_i^T \mathbf{g}_i < 0$. Using standard linear algebra and (7.37), we get

$$\mathbf{x}_i^k - P_{\hat{\Omega}_i}(\mathbf{x}_i^k - \alpha \mathbf{g}_i) = \alpha \hat{\mathbf{g}}_i = \alpha \mathbf{g}_i - \alpha (\mathbf{n}_i^T \mathbf{g}_i)^- \mathbf{n}_i = \alpha \mathbf{g}_i - \alpha (\hat{\mathbf{n}}_i^T \mathbf{g}_i)^- \hat{\mathbf{n}}_i .$$

Thus

$$\alpha \hat{\mathbf{g}}_i = \alpha \mathbf{g}_i^P = \alpha \hat{\mathbf{g}}_i^P .$$

If $i \in \mathscr{F}(\mathbf{x}^k)$ or $\mathbf{n}_i^T \mathbf{g}_i \geq 0$, then obviously

$$\mathbf{g}_i^P = \hat{\mathbf{g}}_i^P = \hat{\mathbf{g}}_i = \mathbf{g}_i . \qquad \square$$

We shall also use the following lemma on the three subsequent iterations that is due to Kučera [1].

Lemma 7.9 *Let $\boldsymbol{\xi}^0$, $\boldsymbol{\xi}^1$, and $\boldsymbol{\xi}^2$ belong to $\hat{\Omega}$ and satisfy*

$$f\left(\boldsymbol{\xi}^2\right) - f(\widehat{\boldsymbol{\xi}}) \leq \eta \left(f(\boldsymbol{\xi}^1) - f(\widehat{\boldsymbol{\xi}})\right) \leq \eta^2 \left(f(\boldsymbol{\xi}^0) - f(\widehat{\boldsymbol{\xi}})\right), \tag{7.39}$$

where $\widehat{\boldsymbol{\xi}}$ denotes the solution of (7.36) and $\eta \in (0, 1)$.

Then

$$f(\boldsymbol{\xi}^1) - f(\boldsymbol{\xi}^2) \leq \frac{1+\eta}{1-\eta} \eta \left(f(\boldsymbol{\xi}^0) - f(\boldsymbol{\xi}^1)\right).$$

Proof We shall repeatedly apply (7.39). As

$$\begin{aligned} f(\boldsymbol{\xi}^0) - f(\boldsymbol{\xi}^1) &= \left(f(\boldsymbol{\xi}^0) - f(\widehat{\boldsymbol{\xi}})\right) - \left(f(\boldsymbol{\xi}^1) - f(\widehat{\boldsymbol{\xi}})\right) \\ &\geq (1-\eta) \left(f(\boldsymbol{\xi}^0) - f(\widehat{\boldsymbol{\xi}})\right) \\ &\geq \frac{1-\eta}{\eta} \left(f(\boldsymbol{\xi}^1) - f(\widehat{\boldsymbol{\xi}})\right) \end{aligned}$$

and

$$\begin{aligned} f(\boldsymbol{\xi}^1) - f(\boldsymbol{\xi}^2) &= 2 \left(\frac{q(x^k) + f(\boldsymbol{\xi}^0)}{2} - f(\boldsymbol{\xi}^2) \right) \\ &\leq 2 \left(\frac{f(\boldsymbol{\xi}^1) + f(\boldsymbol{\xi}^2)}{2} - f(\widehat{\boldsymbol{\xi}}) \right) \\ &\leq (1+\eta) \left(f(\boldsymbol{\xi}^1) - f(\widehat{\boldsymbol{\xi}})\right), \end{aligned}$$

we get

$$\frac{1}{1+\eta} \left(f(\boldsymbol{\xi}^1) - q(x^{k+1})\right) \leq f(\boldsymbol{\xi}^1) - f(\widehat{\boldsymbol{\xi}}) \leq \frac{\eta}{1-\eta} \left(f(\boldsymbol{\xi}^0) - q(x^k)\right).$$

This completes the proof. $\square$

Now we are ready to prove the main result of this section.

Theorem 7.3 *Let* $\{\mathbf{x}^k\}$ *denote the iterates generated by the MPGP algorithm under the assumptions of Theorem 7.2 with*

$$\alpha \in (0, 2\|\mathsf{A}\|^{-1}) \quad \text{and} \quad \delta \in (0, 1/2).$$

Then for any $k \geq 1$

$$\|\mathbf{g}^P(\mathbf{x}^k)\|^2 \leq a_1 \eta^k \big(f(\mathbf{x}^0) - f(\widehat{\mathbf{x}})\big), \quad a_1 = \frac{2\big(1+\eta\big)}{\widehat{\alpha}\big(1-\eta\big)}, \tag{7.40}$$

where $\widehat{\alpha}$ *and* $\eta = \eta(\delta, \alpha)$ *are defined in Theorem 7.2.*

Proof As we have mentioned above, our main tool is the observation that given $k \geq 1$, we can consider the iterates $\mathbf{x}^{k-1}$ and $\mathbf{x}^k$ as initial iterates for auxiliary problem (7.36).

Let us first show that

$$f(\mathbf{x}^k) - f(\widehat{\boldsymbol{\xi}}) \leq \eta\big(f(\mathbf{x}^{k-1}) - f(\widehat{\boldsymbol{\xi}})\big), \tag{7.41}$$

where $\widehat{\boldsymbol{\xi}}$ denotes a unique solution of (7.36). We shall consider separately two steps that can generate $\mathbf{x}^k$.

If $\mathbf{x}^k$ is generated by the conjugate gradient step for problem (7.1), then

$$\mathbf{x}_i^k = \mathbf{x}_i^{k-1} \quad \text{for} \quad i \in \mathscr{A}(\mathbf{x}^{k-1}).$$

Noticing that

$$\hat{\mathscr{A}}^k(\mathbf{x}) \subseteq \mathscr{A}(\mathbf{x})$$

for any $\mathbf{x} \in \Omega_S$, we get

$$\|\hat{\boldsymbol{\beta}}(\mathbf{x}^{k-1})\| \leq \|\boldsymbol{\beta}(\mathbf{x}^{k-1})\| \quad \text{and} \quad \|\hat{\boldsymbol{\varphi}}(\mathbf{x}^{k-1})\| \geq \|\boldsymbol{\varphi}(\mathbf{x}^{k-1})\|.$$

It follows that $\mathbf{x}^{k-1}$ is proportional also as an iterate for the solution of problem (7.36) and (7.41) holds true by Theorem 7.2.

To prove (7.41) for $\mathbf{x}^k$ generated by the gradient projection step, notice that $\hat{\Omega}^k$ is defined in such a way that

$$\mathbf{x}^k = P_{\Omega_S}\big(\mathbf{x}^{k-1} - \alpha \mathbf{g}(\mathbf{x}^{k-1})\big) = P_{\hat{\Omega}^k}\big(\mathbf{x}^{k-1} - \alpha \mathbf{g}(\mathbf{x}^{k-1})\big).$$

We conclude that (7.41) holds true.

Let us define

$$\boldsymbol{\xi}^0 = \mathbf{x}^{k-1}, \ \boldsymbol{\xi}^1 = \mathbf{x}^k, \ \text{ and } \ \boldsymbol{\xi}^2 = P_{\hat{\Omega}^k}\big(\boldsymbol{\xi}^1 - \alpha \mathbf{g}(\boldsymbol{\xi}^1)\big).$$

Using Theorem 7.2 and (7.41), we get

$$\left(f(\boldsymbol{\xi}^2) - f(\widehat{\boldsymbol{\xi}})\right) \le \eta(\delta,\alpha)\left(f(\boldsymbol{\xi}^1) - f(\widehat{\boldsymbol{\xi}})\right) \le \eta(\delta,\alpha)^2\left(f(\boldsymbol{\xi}^0) - f(\widehat{\boldsymbol{\xi}})\right). \quad (7.42)$$

It follows that the assumptions of Lemma 7.9 are satisfied with $\eta = \eta(\delta,\alpha)$ and

$$\begin{aligned} f(\boldsymbol{\xi}^1) - f(\boldsymbol{\xi}^2) &\le \frac{1+\eta}{1-\eta}\eta\left(f(\boldsymbol{\xi}^0) - f(\boldsymbol{\xi}^1)\right) \\ &= \frac{1+\eta}{1-\eta}\eta\left(f(\mathbf{x}^{k-1}) - f(\mathbf{x}^k)\right) \\ &\le \frac{1+\eta}{1-\eta}\eta\left(f(\mathbf{x}^{k-1}) - f(\widehat{\mathbf{x}})\right) \\ &\le \frac{1+\eta}{1-\eta}\eta^k\left(f(\mathbf{x}^0) - f(\widehat{\mathbf{x}})\right). \end{aligned}$$

Finally, using Lemma 7.8, relations (7.7), and simple manipulations, we get

$$\begin{aligned} f\left(\boldsymbol{\xi}^1\right) - f\left(\boldsymbol{\xi}^2\right) &= f\left(\boldsymbol{\xi}^1\right) - f\left(P_{\hat{\Omega}}\left(\boldsymbol{\xi}^1 - \alpha\mathbf{g}(\boldsymbol{\xi}^1)\right)\right) \\ &= \alpha\hat{\mathbf{g}}_\alpha^T(\boldsymbol{\xi}^1)\mathbf{g}(\boldsymbol{\xi}^1) - \frac{\alpha^2}{2}\hat{\mathbf{g}}_\alpha^T(\boldsymbol{\xi}^1)\mathsf{A}\hat{\mathbf{g}}_\alpha(\boldsymbol{\xi}^1) \\ &\ge \alpha\hat{\mathbf{g}}_\alpha^T(\boldsymbol{\xi}^1)\mathbf{g}(\boldsymbol{\xi}^1) - \frac{\alpha^2}{2}\|\mathsf{A}\|\|\hat{\mathbf{g}}_\alpha(\boldsymbol{\xi}^1)\|^2 \\ &= (\alpha - \frac{\alpha^2}{2}\|\mathsf{A}\|)\,\hat{\mathbf{g}}_\alpha^T(\boldsymbol{\xi}^1)\mathbf{g}(\boldsymbol{\xi}^1) \\ &= \frac{1}{2}\|\mathsf{A}\|\alpha(2\|\mathsf{A}\|^{-1} - \alpha)\,\hat{\mathbf{g}}_\alpha^T(\boldsymbol{\xi}^1)\mathbf{g}(\boldsymbol{\xi}^1) \\ &\ge \frac{\widehat{\alpha}}{2}\hat{\mathbf{g}}_\alpha^T(\boldsymbol{\xi}^1)\mathbf{g}(\boldsymbol{\xi}^1) = \frac{\widehat{\alpha}}{2}\left(\hat{\mathbf{g}}^P(\boldsymbol{\xi}^1)\right)^T\mathbf{g}(\boldsymbol{\xi}^1) \\ &= \frac{\widehat{\alpha}}{2}\|\hat{\mathbf{g}}^P(\boldsymbol{\xi}^1)\|^2 = \frac{\widehat{\alpha}}{2}\|\mathbf{g}^P(\boldsymbol{\xi}^1)\|^2 = \frac{\widehat{\alpha}}{2}\|\mathbf{g}^P(\mathbf{x}^k)\|^2. \end{aligned}$$

To verify the last inequality, consider $\alpha \in (0, \|\mathsf{A}\|^{-1}]$ and $\alpha \in (\|\mathsf{A}\|^{-1}, 2\|\mathsf{A}\|^{-1})$ separately. Putting the last terms of the above chains of relations together, we get (7.40). □

7.6 Implementation

In this section, we describe Algorithm 7.1 in the form that is convenient for implementation. We include also some modifications that may be used to improve its performance.

7.6.1 Projection Step with Feasible Half-Step

To improve the efficiency of the projection step, we can use the trial conjugate gradient direction $\mathbf{p}^k$ which is generated before the projection step is invoked. We propose to generate first

$$\mathbf{x}^{k+\frac{1}{2}} = \mathbf{x}^k - \alpha_f \mathbf{p}^k \quad \text{and} \quad \mathbf{g}^{k+\frac{1}{2}} = \mathbf{g}^k - \alpha_f \mathsf{A}\mathbf{p}^k,$$

where the feasible step length α_f for $\mathbf{p}^k$ is defined by

$$\alpha_f = \max\{\alpha : \ \mathbf{x}^k - \alpha \mathbf{p}^k \in \Omega_S\}, \tag{7.43}$$

and then define

$$\mathbf{x}^{k+1} = P_{\Omega_S}\left(\mathbf{x}^{k+\frac{1}{2}} - \alpha \mathbf{g}(\mathbf{x}^{k+\frac{1}{2}})\right).$$

The half-step is illustrated in Fig. 7.3. Such modification does not require any additional matrix–vector multiplication and estimate (7.32) remains valid as

$$f(\mathbf{x}^{k+\frac{1}{2}}) - f(\mathbf{x}^k) \le 0$$

and

$$\begin{aligned} f(\mathbf{x}^{k+1}) - f(\widehat{\mathbf{x}}) &\le \eta_\Gamma \left((f(\mathbf{x}^{k+\frac{1}{2}}) - f(\mathbf{x}^k)) + f(\mathbf{x}^k) - f(\widehat{\mathbf{x}}) \right) \\ &\le \eta_\Gamma \left(f(\mathbf{x}^k) - f(\widehat{\mathbf{x}}) \right). \end{aligned}$$

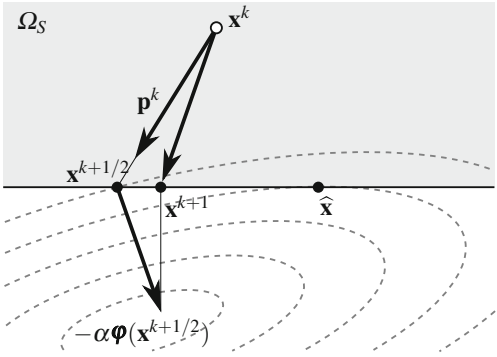

Fig. 7.3 Feasible half-step

Since our analysis is based on the worst-case analysis, the implementation of the feasible half-step does not result in improving the error bounds.

7.6.2 MPGP Algorithm in More Detail

Now we are ready to give the details of implementation of the MPGP algorithm which was briefly described in a form suitable for analysis as Algorithm 7.1. To preserve readability, we do not distinguish the generations of variables by indices unless it is convenient for further reference.

Algorithm 7.2 MPGP with a feasible half-step.

Given a symmetric positive definite matrix $\mathsf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, *and* Ω_S.

Step 0. { *Initialization of parameters.*}

Choose $\mathbf{x}^0 \in \Omega_S$, $\alpha \in (0, 2\|\mathsf{A}\|^{-1})$, $\Gamma > 0$, *and the relative stopping tolerance* $\varepsilon > 0$. *Set* $k = 0$, $\mathbf{g} = \mathsf{A}\mathbf{x}^0 - \mathbf{b}$, $\mathbf{p} = \boldsymbol{\varphi}(\mathbf{x}^0)$.

while $\left\|\widetilde{\mathbf{g}}^P_\alpha\left(\mathbf{x}^k\right)\right\|$ *is not small*

if $\|\boldsymbol{\beta}(\mathbf{x}^k)\| \le \Gamma\|\boldsymbol{\varphi}(\mathbf{x}^k)\|$

Step 1. {*Proportional* $\mathbf{x}^k$. *Trial conjugate gradient step.*}

$\alpha_{cg} = \mathbf{g}^T\mathbf{p}/\mathbf{p}^T\mathsf{A}\mathbf{p}$

$\alpha_f = \max\left\{\alpha : \mathbf{x}^k - \alpha\mathbf{p} \in \Omega_S\right\}$

if $\alpha_{cg} \le \alpha_f$

Step 2. { *Conjugate gradient step.*}

$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_{cg}\mathbf{p}$, $\mathbf{g} = \mathbf{g} - \alpha_{cg}\mathsf{A}\mathbf{p}$

$\gamma = \boldsymbol{\varphi}\left(\mathbf{x}^{k+1}\right)^T\mathsf{A}\mathbf{p}/\mathbf{p}^T\mathsf{A}\mathbf{p}$, $\mathbf{p} = \boldsymbol{\varphi}\left(\mathbf{x}^{k+1}\right) - \gamma\mathbf{p}$

else

Step 3. {*Gradient projection step with halfstep.*}

$\mathbf{x}^{k+\frac{1}{2}} = \mathbf{x}^k - \alpha_f\mathbf{p}$, $\mathbf{g} = \mathbf{g} - \alpha_f\mathsf{A}\mathbf{p}$

$\mathbf{x}^{k+1} = P_{\Omega_S}\left(\mathbf{x}^{k+\frac{1}{2}} - \overline{\alpha}\mathbf{g}\right)$

$\mathbf{g} = \mathsf{A}\mathbf{x}^{k+1} - \mathbf{b}$, $\mathbf{p} = \boldsymbol{\varphi}\left(\mathbf{x}^{k+1}\right)$

end if

else

Step 4. {*Gradient projection step.*}

$\mathbf{x}^{k+1} = P_{\Omega_S}\left(\mathbf{x}^k - \alpha\mathbf{g}\right)$

$\mathbf{g} = \mathsf{A}\mathbf{x}^{k+1} - \mathbf{b}$, $\mathbf{p} = \boldsymbol{\varphi}\left(\mathbf{x}^{k+1}\right)$

end if

$k = k + 1$

end while

Step 5. {*Return (possibly inexact) solution.*}

$\widetilde{\mathbf{x}} = \mathbf{x}^k$

Our experience indicates that the performance of MPGP is not sensitive to Γ as long as $\Gamma \approx 1$. Since $\Gamma = 1$ minimizes the upper bound on the rate of convergence and guarantees that the CG steps reduce directly the larger of the two components of the projected gradient, we can expect good efficiency with this value. Recall that $\Gamma = 1$ corresponds to $\delta = 1/4$.

The choice of α requires an estimate of $\|\mathsf{A}\|$. If we cannot exploit a specific structure of A, then we can carry out a few, e.g., five, iterations of the following power method.

Algorithm 7.3 Power method for the estimate of $\|\mathsf{A}\|$.

Given a symmetric positive definite matrix $\mathsf{A} \in \mathbb{R}^{n\times n}$, *returns* $A \approx \|\mathbf{A}\|$. *Choose* $\mathbf{x} \in \mathbb{R}^n$ *such that* $\mathbf{x} \neq \mathbf{o}$, $n_{\rm it} \geq 1$

for $i = 1, 2, \ldots, n_{\rm it}$

$\mathbf{y} = \mathsf{A}\mathbf{x}$, $\mathbf{x} = \|\mathbf{y}\|^{-1}\mathbf{y}$

end for

$A = \|\mathsf{A}\mathbf{x}\|$

Alternatively, we can use the Lanczos method (see, e.g., Golub and van Loan [2]). We can conveniently enhance the Lanczos method into the first conjugate gradient loop of the MPRGP algorithm by defining

$$\mathbf{q}_i = \|\boldsymbol{\varphi}(\mathbf{x}^{s+i})\|^{-1}\boldsymbol{\varphi}(\mathbf{x}^{s+i}), \quad i = 0, \ldots, p,$$

where $\boldsymbol{\varphi}(\mathbf{x}^s)$ and $\boldsymbol{\varphi}(\mathbf{x}^{s+i})$ are free gradients at the initial and the ith iterate in one CG loop, respectively. Then we can estimate $\|\mathsf{A}\|$ by applying a suitable method for evaluation of the norm of the tridiagonal matrix [2]

$$\mathsf{T} = \mathsf{Q}^T\mathsf{A}\mathsf{Q}, \quad \mathsf{Q} = [\mathbf{q}_0, \ldots, \mathbf{q}_p].$$

Though these methods typically give only a lower bound A on the norm of $\|\mathsf{A}\|$, the choice like $\alpha = 1.8A^{-1}$ is often sufficient in practice. The decrease of f can be achieved more reliably by initializing $\alpha \geq 2(\mathbf{b}^T\mathsf{A}\mathbf{b})^{-1}\|\mathbf{b}\|^2$ and by inserting the following piece of code into the expansion step:

Algorithm 7.4 Modification of the steplength of the expansion step.

A piece of code to be inserted at the end of the expansion step of Algorithm 8.2.

if $f\left(P_{\Omega_B}(\mathbf{x}^{k+1})\right) > f(\mathbf{x}^k)$

$\alpha = \alpha/2$ *and repeat the expansion step*

end if

The modified algorithm can outperform that with $\alpha = \|\mathsf{A}\|^{-1}$ as longer steps in the early stage of computations can be effective for fast identification of the active set of the solution. We observed a good performance with α close to, but not greater than $2\|\mathsf{A}\|^{-1}$, near α_E^{opt} which minimizes the coefficient η_E of the Euclidean contraction (6.8).

7.7 Comments and References

The MPGP algorithm presented in this section is a variant of the algorithms which combine conjugate gradients with Euclidean projections. Such algorithms were developed first for solving the bound constrained QP problems. See Sect. 8.7 for more information.

It seems that the first algorithm of this type for separable QCQP problems was presented by Kučera [3], who found the way how to adapt the MPRGP algorithm ([4], see the next chapter), originally proposed for the solution of bound constrained QP problems, for solving more general problems. He proved the R-linear convergence of his KPRGP algorithm for the step length $\alpha \in (0, \|\mathsf{A}\|^{-1}]$. Later he also proved the R-linear convergence of projected gradients for the step length $\alpha \in (0, \|\mathsf{A}\|^{-1}]$ [1].

The MPGP algorithm presented here appeared in Dostál and Kozubek [5]. The estimate of the rate of convergence presented in Theorem 7.2 is slightly better than that in [1] and guarantees the R-linear convergence of the projected gradient for $\alpha \in (0, 2\|\mathsf{A}\|^{-1}]$. The proof uses the estimates due to Bouchala, Dostál, and Vodstrčil [6, 7].

Here we provided the analysis of a monotonically decreasing algorithm with the rate of convergence in bounds on the spectrum of A. However, we observed that its performance can be sometime improved using some heuristic modifications, in particular those using longer step length or unfeasible steps, such as the conjugate gradient or Barzialai–Borwein step length for the solution of a related minimization problem in free variables or the heuristics proposed in [8]. See also Sect. 6.6.

The performance of MPGP can also be improved by *preconditioning*. We have postponed the description of preconditioning to Sect. 8.6 in order to exploit the simplified setting of the bound constrained QP problem. The preconditioning in face improves the solution of auxiliary unconstrained problems, while the preconditioning by a conjugate projector improves the efficiency of all steps, including the nonlinear ones. The effective preconditioners are problem dependent. The preconditioners suitable for the solution of contact problems which comply with the FETI methodology are in Chap. 16. The preconditioning by the conjugate projector (deflation) is described in Sect. 13.6 in the context of transient contact problems.

References

1. Dostál, Z., Kučera, R.: An optimal algorithm for minimization of quadratic functions with bounded spectrum subject to separable convex inequality and linear equality constraints. SIAM J. Optim. **20**(6), 2913–2938 (2010)
2. Golub, G.H., Van Loan, C.F.: Matrix Computations, 2nd edn. Johns Hopkins University Press, Baltimore (1989)
3. Kučera, R.: Convergence rate of an optimization algorithm for minimizing quadratic functions with separable convex constraints. SIAM J. Optim. **19**, 846–862 (2008)
4. Dostál, Z., Schöberl, J.: Minimizing quadratic functions over non-negative cone with the rate of convergence and finite termination. Comput. Optim. Appl. **30**(1), 23–44 (2005)

5. Dostál, Z., Kozubek, T.: An optimal algorithm with superrelaxation for minimization of a quadratic function subject to separable constraints with applications. Math. Program. Ser. A **135**, 195–220 (2012)
6. Bouchala, J., Dostál, Z., Vodstrčil, P.: Separable spherical constraints and the decrease of a quadratic function in the gradient projection. J. Optim. Theory Appl. **157**, 132–140 (2013)
7. Bouchala, J., Dostál, Z., Kozubek, T., Pospíšil, L., Vodstrčil, P.: On the solution of convex QPQC problems with elliptic and other separable constraints. Appl. Math. Comput. **247**(15), 848–864 (2014)
8. Dostál, Z.: Box constrained quadratic programming with proportioning and projections. SIAM J. Optim. **7**(3), 871–887 (1997)

Chapter 8
MPRGP for Bound-Constrained QP

We shall now be concerned with a special case of separable problem (7.1), the *bound-constrained problem* to find

$$\min_{\mathbf{x}\in\Omega_B} f(\mathbf{x}) \tag{8.1}$$

with

$$\Omega_B = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq \ell\}, \quad f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathsf{A}\mathbf{x} - \mathbf{x}^T\mathbf{b},$$

ℓ and $\mathbf{b}$ given column n-vectors, and A an $n \times n$ SPD matrix. To include the possibility that not all components of $\mathbf{x}$ are constrained, we admit $\ell_i = -\infty$. The problem (8.1) appears in the dual formulation of both static and dynamic contact problems without friction.

There are two specific features of (8.1) that are not explicitly exploited by the MPGP algorithm of Chap. 7, namely the possibility to move arbitrarily in the direction opposite to the chopped gradient and a simple observation that knowing the active constraints of the solution amounts to knowing the corresponding components of the solution. Here, we present a modification of MPGP that is able to exploit these features. The modified algorithm is a variant of the active set strategy that we coined *MPRGP* (Modified Proportioning with Reduced Gradient Projections). The algorithm uses the conjugate gradients to solve the auxiliary unconstrained problems with the precision controlled by the norm of the dominating component of the projected gradient. The fixed steplength reduced gradient projections and the optimal steplength chopped gradient steps are used to expand and reduce the active set, respectively.

It turns out that MPRGP has not only the R-linear rate of convergence in terms of the extreme eigenvalues of the Hessian matrix as MPGP of the previous chapter but also the finite termination property, even in the case that the solution is dual degenerate. We consider the finite termination property important, as it indicates that the algorithm does not suffer from undesirable oscillations often attributed to the active set-based algorithms and thus can better exploit the superconvergence properties of the conjugate gradient method for linear problems.

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3_8

8.1 Specific Form of KKT Conditions

Let us introduce special notations that enable us to simplify the form of the projected gradient (7.6) and of the KKT conditions which read

$$\mathbf{g}^P(\mathbf{x}) = \mathbf{o}. \tag{8.2}$$

The KKT conditions at $\mathbf{x} \in \Omega_B$ determine three subsets of the set $\mathcal{N} = \{1, \ldots, n\}$ of all indices. The set of all indices for which $x_i = \ell_i$ is called an *active set* of $\mathbf{x}$. We denote it by $\mathcal{A}(\mathbf{x})$, so

$$\mathcal{A}(\mathbf{x}) = \{i \in \mathcal{N} : \ x_i = \ell_i\}.$$

Its complement

$$\mathcal{F}(\mathbf{x}) = \{i \in \mathcal{N} : \ x_i \neq \ell_i\}$$

and subsets

$$\mathcal{B}(\mathbf{x}) = \{i \in \mathcal{N} : \ x_i = \ell_i \text{ and } g_i > 0\}, \quad \mathcal{B}_0(\mathbf{x}) = \{i \in \mathcal{N} : \ x_i = \ell_i \text{ and } g_i \geq 0\}$$

are called a *free set*, a *binding set*, and a *weakly binding set*, respectively.

Using the subsets of $\mathcal{N}$, we can decompose $\mathbf{g}^P(\mathbf{x})$ into the *free gradient* $\boldsymbol{\varphi}$ and the *chopped gradient* (Fig. 8.1) $\boldsymbol{\beta}$ that are defined by

$$\begin{aligned} &\varphi_i(\mathbf{x}) = g_i(\mathbf{x}) \text{ for } i \in \mathcal{F}(\mathbf{x}), \quad &&\varphi_i(\mathbf{x}) = 0 \text{ for } i \in \mathcal{A}(\mathbf{x}), \\ &\beta_i(\mathbf{x}) = 0 \text{ for } i \in \mathcal{F}(\mathbf{x}), &&\beta_i(\mathbf{x}) = g_i^-(\mathbf{x}) \text{ for } i \in \mathcal{A}(\mathbf{x}), \end{aligned}$$

where we have used the notation $g_i^- = \min\{g_i, 0\}$. Thus

$$\mathbf{g}^P(\mathbf{x}) = \boldsymbol{\varphi}(\mathbf{x}) + \boldsymbol{\beta}(\mathbf{x}).$$

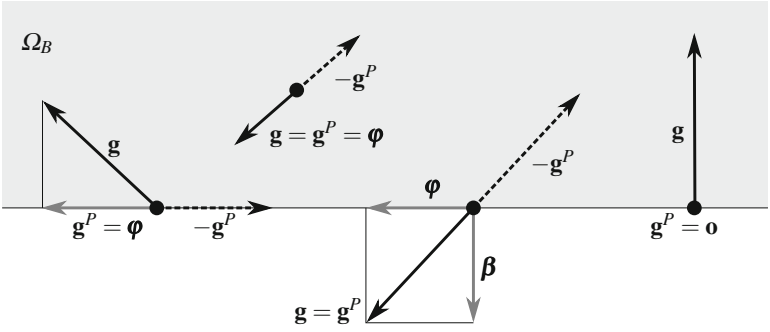

Fig. 8.1 Gradient splitting

8.2 MPRGP Algorithm

Let us first recall that if the indices of the active constraints of the solution are known, then the corresponding components are known. There is a simple result which shows that if the norm of the chopped gradient is sufficiently larger than the norm of the free gradient, then it is possible to recognize some indices of active constraints that do not belong to the solution active set and use the components of the chopped gradient to reduce the active set [1, Lemma 5.4]. Notice that the latter can be done much more efficiently with bound constraints then with more general constraints, as any step in the direction opposite to the chopped gradient is feasible.

MPRGP enhances these observations by replacing the gradient projection step of MPGP which *changes* the active set by the *free gradient projection* with a fixed steplength which *expands* the active set. The modified algorithm has been proved to preserve the R-linear rate of convergence of the cost function and to *enjoy the finite termination property even for QP problems* with dual degenerate solution.

The MPRGP algorithm exploits a user-defined constant $\Gamma > 0$, a test which is used to decide when to leave the face, and three types of steps.

The *conjugate gradient step*, defined by

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_{cg}\mathbf{p}^{k+1}, \tag{8.3}$$

is used in the same way as in the MPGP algorithm introduced in Sect. 7.3.

The active set is expanded by the *expansion step* defined by the free gradient projection

$$\mathbf{x}^{k+1} = P_{\Omega_B}\big(\mathbf{x}^k - \alpha\boldsymbol{\varphi}(\mathbf{x}^k)\big) = \max\{\ell, \mathbf{x}^k - \alpha\boldsymbol{\varphi}(\mathbf{x}^k)\} \tag{8.4}$$

with a fixed steplength α. To describe it in the form suitable for analysis, let us recall that, for any $\mathbf{x} \in \Omega_B$ and $\alpha > 0$, the *reduced free gradient* $\widetilde{\boldsymbol{\varphi}}_\alpha(\mathbf{x})$ is defined by the entries

$$\widetilde{\varphi}_i = \widetilde{\varphi}_i(\mathbf{x}, \alpha) = \min\{(x_i - \ell_i)/\alpha, \varphi_i\}, \quad i \in \mathcal{N} = \{1, \ldots, n\}, \tag{8.5}$$

so that

$$P_{\Omega_B}\big(\mathbf{x} - \alpha\boldsymbol{\varphi}(\mathbf{x})\big) = \mathbf{x} - \alpha\widetilde{\boldsymbol{\varphi}}_\alpha(\mathbf{x}). \tag{8.6}$$

Using this notation, we can write also

$$P_{\Omega_B}\big(\mathbf{x} - \alpha\mathbf{g}(\mathbf{x})\big) = \mathbf{x} - \alpha\big(\widetilde{\boldsymbol{\varphi}}_\alpha(\mathbf{x}) + \boldsymbol{\beta}(\mathbf{x})\big). \tag{8.7}$$

If the steplength is equal to α and the inequality

$$||\boldsymbol{\beta}(\mathbf{x}^k)||^2 \le \Gamma^2\widetilde{\boldsymbol{\varphi}}_\alpha(\mathbf{x}^k)^T\boldsymbol{\varphi}(\mathbf{x}^k) \tag{8.8}$$

holds, then we call the iterate $\mathbf{x}^k$ *strictly proportional*. Test (8.8) is used to decide which components of the projected gradient $\mathbf{g}^P(\mathbf{x}^k)$ should be reduced in the next

step. Notice that the right-hand side of (8.8) blends the information about the free gradient and its part that can be used in the gradient projection step.

It is possible to replace the free gradient by some other direction, e.g., $\mathbf{g}^-$. We have made some experiments, but have not found much difference.

The *proportioning step* is defined by

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_{cg}\boldsymbol{\beta}(\mathbf{x}^k) \tag{8.9}$$

with the steplength

$$\alpha_{cg} = \arg\min_{\alpha>0} f\left(\mathbf{x}^k - \alpha\boldsymbol{\beta}(\mathbf{x}^k)\right).$$

It has been shown in Sect. 5.1 that the CG steplength α_{cg} that minimizes $f(\mathbf{x}-\alpha\mathbf{d})$ for a given $\mathbf{d}$ and $\mathbf{x}$ can be evaluated using the gradient $\mathbf{g} = \mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})$ at $\mathbf{x}$ by

$$\alpha_{cg} = \alpha_{cg}(\mathbf{d}) = \mathbf{d}^T\mathbf{g}/\mathbf{d}^T\mathsf{A}\mathbf{d}. \tag{8.10}$$

The purpose of the proportioning step is to remove the indices of the components of the gradient $\mathbf{g}$ that violate the KKT conditions from the working set and to move far from the bounds. Note that if $\mathbf{x}^k \in \Omega_B$, then

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_{cg}\boldsymbol{\beta}(\mathbf{x}^k) \in \Omega_B.$$

Now we are ready to define the algorithm in the form that is convenient for analysis, postponing the discussion about implementation to the next section.

Algorithm 8.1 Modified proportioning with reduced gradient projections (MPRGP schema).

Given an SPD matrix $\mathsf{A} \in \mathbb{R}^{n\times n}$ *and* n*-vectors* $\mathbf{b}$, ℓ.
Choose $\mathbf{x}^0 \in \Omega_B$, $\alpha \in \left(0, 2\|\mathsf{A}\|^{-1}\right)$, *and* $\Gamma > 0$. *Set* $k = 0$.
For $k \geq 0$ *and* $\mathbf{x}^k$ *known, choose* $\mathbf{x}^{k+1}$ *by the following rules:*

(i) *If* $\mathbf{g}^P(\mathbf{x}^k) = \mathbf{o}$, *set* $\mathbf{x}^{k+1} = \mathbf{x}^k$.

(ii) *If* $\mathbf{x}^k$ *is strictly proportional and* $\mathbf{g}^P(\mathbf{x}^k) \neq \mathbf{o}$, *try to generate* $\mathbf{x}^{k+1}$ *by the conjugate gradient step. If* $\mathbf{x}^{k+1} \in \Omega_B$, *then accept it, else generate* $\mathbf{x}^{k+1}$ *by the expansion step.*

(iii) *If* $\mathbf{x}^k$ *is not strictly proportional, define* $\mathbf{x}^{k+1}$ *by proportioning.*

We call our algorithm modified proportioning to distinguish it from earlier algorithms introduced independently by Friedlander and Martínez with their collaborators [2] and Dostál [3].

8.3 Rate of Convergence

The result on the rate of convergence of both the iterates generated by MPRGP and the projected gradient reads as follows.

Theorem 8.1 *Let* $\{\mathbf{x}^k\}$ *be generated by Algorithm 8.1 with* $\mathbf{x}^0 \in \Omega_B$, $\Gamma > 0$, *and* $\alpha \in (0, 2\|\mathsf{A}\|^{-1})$. *Let* $\widehat{\mathbf{x}}$ *and* $\lambda_{\min}$ *denote a unique solution of (8.1) and the smallest eigenvalue of* A, *respectively.*

Then for any $k \geq 1$

$$f(\mathbf{x}^{k+1}) - f(\widehat{\mathbf{x}}) \leq \eta_\Gamma \left(f(\mathbf{x}^k) - f(\widehat{\mathbf{x}})\right), \tag{8.11}$$

$$\|\mathbf{g}^P(\mathbf{x}^{k+1})\|^2 \leq a_1 \eta_\Gamma^k \left(f(\mathbf{x}^0) - f(\widehat{\mathbf{x}})\right), \tag{8.12}$$

$$\|\mathbf{x}^k - \widehat{\mathbf{x}}\|_{\mathsf{A}}^2 \leq 2\eta_\Gamma^k \left(f(\mathbf{x}^0) - f(\widehat{\mathbf{x}})\right), \tag{8.13}$$

where

$$\eta_\Gamma = 1 - \frac{\widehat{\alpha}\lambda_{\min}}{\vartheta + \vartheta\widehat{\Gamma}^2}, \qquad \widehat{\Gamma} = \max\{\Gamma, \Gamma^{-1}\}, \tag{8.14}$$

$$\vartheta = 2\max\{\alpha\|\mathsf{A}\|, 1\} \leq 4, \quad \widehat{\alpha} = \min\{\alpha, 2\|\mathsf{A}\|^{-1} - \alpha\}, \tag{8.15}$$

and

$$a_1 = \frac{1 + \eta_\Gamma}{2\widehat{\alpha}(1 - \eta_\Gamma)}. \tag{8.16}$$

Proof The proof of (8.11) is technical and may be found in Domorádová, Dostál, and Sadowská [4] or Dostál [1]. The proof of (8.12) is a simplified version of the proof of Theorem 7.3 based on estimate (8.11). Notice that Lemma 7.9 on three iterates can be applied directly to $\mathbf{x}^{k-1}$, $\mathbf{x}^k$, and $\mathbf{x}^{k+1}$. □

Theorem 8.1 gives the best bound on the rate of convergence for $\Gamma = \widehat{\Gamma} = 1$ in agreement with the heuristics that we should leave the face when the chopped gradient dominates the violation of the Karush–Kuhn–Tucker conditions. The formula for the best bound η_Γ^{opt} which corresponds to $\Gamma = 1$ and $\alpha = \|\mathsf{A}\|^{-1}$ reads

$$\eta_\Gamma^{opt} = 1 - \kappa(\mathsf{A})^{-1}/4, \tag{8.17}$$

where $\kappa(\mathsf{A})$ denotes the spectral condition number of A.

The bound on the rate of convergence of the projected gradient given by (8.12) is rather poor. The reason is that it has been obtained by the worst case analysis of a general couple of consecutive iterations and does not reflect the structure of a longer chain of the same type of iterations. Recall that Fig. 7.2 shows that no bound on $\mathbf{g}^P(\mathbf{x}^k)$ can be obtained by the analysis of a single iteration!

8.4 Identification Lemma and Finite Termination

Let us consider the conditions which guarantee that the MPRGP algorithm finds the solution $\widehat{\mathbf{x}}$ of (8.1) in a finite number of steps. Recall that such algorithm is more likely to generate longer sequences of the conjugate gradient iterations. In this case the reduction of the cost function values is bounded by the "global" estimate (5.15), and finally switches to the conjugate gradient method, so that it can exploit its nice self-acceleration property [5]. It is difficult to enhance these characteristics of the algorithm into the rate of convergence as they cannot be obtained by the analysis of just one step of the method.

We first examine the finite termination of Algorithm 8.1 in a simpler case when the solution $\widehat{\mathbf{x}}$ of (8.1) is *not dual degenerate*, i.e., the vector of Lagrange multipliers $\widehat{\lambda}$ of the solution satisfies the *strict complementarity condition* $\widehat{\lambda}_i > 0$ for $i \in \mathscr{A}(\widehat{\mathbf{x}})$. The proof is based on simple geometrical observations and the arguments proposed by Moré and Toraldo [6]. For example, it is easy to see that the free sets of the iterates $\mathbf{x}^k$ soon contain the free set of the solution $\widehat{\mathbf{x}}$. The formal analysis of such observations is a subject of the following identification lemma.

Lemma 8.1 *Let* $\{\mathbf{x}^k\}$ *be generated by Algorithm 8.1 with* $\mathbf{x}^0 \in \Omega_B$, $\Gamma > 0$, *and* $\alpha \in (0, 2\|\mathsf{A}\|^{-1}]$. *Then there is* k_0 *such that for* $k \geq k_0$

$$\mathscr{F}(\widehat{\mathbf{x}}) \subseteq \mathscr{F}(\mathbf{x}^k), \quad \mathscr{F}(\widehat{\mathbf{x}}) \subseteq \mathscr{F}(\mathbf{x}^k - \alpha\widetilde{\boldsymbol{\varphi}}(\mathbf{x}^k)), \quad \text{and} \quad \mathscr{B}(\widehat{\mathbf{x}}) \subseteq \mathscr{B}(\mathbf{x}^k), \tag{8.18}$$

where $\widetilde{\boldsymbol{\varphi}}(\mathbf{x}^k) = \widetilde{\boldsymbol{\varphi}}_\alpha(\mathbf{x}^k)$ *is defined by (8.5).*

Proof Since (8.18) is trivially satisfied when there is $k = k_0$ such that $\mathbf{x}^k = \widehat{\mathbf{x}}$, we shall assume in what follows that $\mathbf{x}^k \neq \widehat{\mathbf{x}}$ for any $k \geq 0$. Let us denote $x_i^k = [\mathbf{x}^k]_i$ and $\widehat{x}_i = [\widehat{\mathbf{x}}^k]_i$, $i = 1, \ldots, n$.

Let us first assume that $\mathscr{F}(\widehat{\mathbf{x}}) \neq \emptyset$ and $\mathscr{B}(\widehat{\mathbf{x}}) \neq \emptyset$, so that we can define

$$\varepsilon = \min\{\widehat{x}_i - \ell_i : i \in \mathscr{F}(\widehat{\mathbf{x}})\} > 0 \quad \text{and} \quad \delta = \min\{g_i(\widehat{\mathbf{x}}) : i \in \mathscr{B}(\widehat{\mathbf{x}})\} > 0.$$

Since $\{\mathbf{x}^k\}$ converges to $\widehat{\mathbf{x}}$ by Theorem 8.1, there is k_0 such that for any $k \geq k_0$

$$g_i(\mathbf{x}^k) \leq \frac{\varepsilon}{4\alpha} \text{ for } i \in \mathscr{F}(\widehat{\mathbf{x}}), \tag{8.19}$$

$$x_i^k \geq \ell_i + \frac{\varepsilon}{2} \text{ for } i \in \mathscr{F}(\widehat{\mathbf{x}}), \tag{8.20}$$

$$x_i^k \leq \ell_i + \frac{\alpha\delta}{8} \text{ for } i \in \mathscr{B}(\widehat{\mathbf{x}}), \tag{8.21}$$

$$g_i(\mathbf{x}^k) \geq \frac{\delta}{2} \text{ for } i \in \mathscr{B}(\widehat{\mathbf{x}}). \tag{8.22}$$

In particular, for $k \geq k_0$, the first inclusion of (8.18) follows from (8.20), while the second inclusion follows from (8.19) and (8.20), as for $i \in \mathscr{F}(\widehat{\mathbf{x}})$

$$x_i^k - \alpha\varphi_i(\mathbf{x}^k) = x_i^k - \alpha g_i(\mathbf{x}^k) \geq \ell_i + \frac{\varepsilon}{2} - \frac{\alpha\varepsilon}{4\alpha} > \ell_i.$$

Let $k \geq k_0$ and observe that, by (8.21) and (8.22), for any $i \in \mathscr{B}(\widehat{\mathbf{x}})$

$$x_i^k - \alpha g_i(\mathbf{x}^k) \leq \ell_i + \frac{\alpha\delta}{8} - \frac{\alpha\delta}{2} < \ell_i,$$

so that if some $\mathbf{x}^{k+1}$ is generated by the expansion step (8.4), $k \geq k_0$, and $i \in \mathscr{B}(\widehat{\mathbf{x}})$, then

$$x_i^{k+1} = \max\{\ell_i, x_i^k - \alpha g_i(\mathbf{x}^k)\} = \ell_i.$$

It follows that if $k \geq k_0$ and $\mathbf{x}^{k+1}$ is generated by the expansion step, then $\mathscr{B}(\mathbf{x}^{k+1}) \supseteq \mathscr{B}(\widehat{\mathbf{x}})$. Moreover, using (8.22) and the definition of Algorithm 8.1, we can directly verify that if $\mathscr{B}(\mathbf{x}^k) \supseteq \mathscr{B}(\widehat{\mathbf{x}})$ and $k \geq k_0$, then also $\mathscr{B}(\mathbf{x}^{k+1}) \supseteq \mathscr{B}(\widehat{\mathbf{x}})$. Thus it remains to prove that there is $s \geq k_0$ such that $\mathbf{x}^s$ is generated by the expansion step.

Let us examine what can happen for $k \geq k_0$. First observe that we can never take the full CG step in the direction $\mathbf{p}^k = \boldsymbol{\varphi}(\mathbf{x}^k)$. The reason is that

$$\alpha_{cg}(\mathbf{p}^k) = \frac{\boldsymbol{\varphi}(\mathbf{x}^k)^T\mathbf{g}(\mathbf{x}^k)}{\boldsymbol{\varphi}(\mathbf{x}^k)^T\mathsf{A}\boldsymbol{\varphi}(\mathbf{x}^k)} = \frac{\|\boldsymbol{\varphi}(\mathbf{x}^k)\|^2}{\boldsymbol{\varphi}(\mathbf{x}^k)^T\mathsf{A}\boldsymbol{\varphi}(\mathbf{x}^k)} \geq \|\mathsf{A}\|^{-1} \geq \frac{\alpha}{2},$$

so that for $i \in \mathscr{F}(\mathbf{x}^k) \cap \mathscr{B}(\widehat{\mathbf{x}})$, by (8.21) and (8.22),

$$x_i^k - \alpha_{cg} p_i^k = x_i^k - \alpha_{cg} g_i(\mathbf{x}^k) \leq x_i^k - \frac{\alpha}{2} g_i(\mathbf{x}^k) \leq \ell_i + \frac{\alpha\delta}{8} - \frac{\alpha\delta}{4} < \ell_i. \tag{8.23}$$

It follows by the definition of Algorithm 8.1 that if $\mathbf{x}^k$, $k \geq k_0$, is generated by the proportioning step, then the following trial conjugate gradient step is not feasible, and $\mathbf{x}^{k+1}$ is necessarily generated by the expansion step.

To complete the proof, observe that Algorithm 8.1 can generate only a finite sequence of consecutive conjugate gradient iterates. Indeed, if there is neither proportioning step nor the expansion step for $k \geq k_0$, then it follows by the finite termination property of the conjugate gradient method that there is $l \leq n$ such that $\boldsymbol{\varphi}(\mathbf{x}^{k_0+l}) = \mathbf{o}$. Thus either $\mathbf{x}^{k_0+l} = \widehat{\mathbf{x}}$ and $\mathscr{B}(\mathbf{x}^k) = \mathscr{B}(\widehat{\mathbf{x}})$ for $k \geq k_0 + l$ by rule (i), or $\mathbf{x}^{k_0+l}$ is not strictly proportional, $\mathbf{x}^{k_0+l+1}$ is generated by the proportioning step, and $\mathbf{x}^{k_0+l+2}$ is generated by the expansion step. This completes the proof, as the cases $\mathscr{F}(\widehat{\mathbf{x}}) = \emptyset$ and $\mathscr{B}(\widehat{\mathbf{x}}) = \emptyset$ can be proved by the analysis of the above arguments. □

Proposition 8.1 *Let $\{\mathbf{x}^k\}$ be generated by Algorithm 8.1 with $\mathbf{x}^0 \in \Omega_B$, $\Gamma > 0$, and $\alpha \in \left(0, 2\|\mathsf{A}\|^{-1}\right)$. Let the solution $\widehat{\mathbf{x}}$ satisfy the condition of strict complementarity, i.e., $\widehat{x}_i = \ell_i$ implies $g_i(\widehat{\mathbf{x}}) > 0$. Then there is $k \geq 0$ such that $\mathbf{x}^k = \widehat{\mathbf{x}}$.*

Proof If $\widehat{\mathbf{x}}$ satisfies the condition of strict complementarity, then $\mathscr{A}(\widehat{\mathbf{x}}) = \mathscr{B}(\widehat{\mathbf{x}})$, and, by Lemma 8.1, there is $k_0 \geq 0$ such that for $k \geq k_0$ we have $\mathscr{F}(\mathbf{x}^k) = \mathscr{F}(\widehat{\mathbf{x}})$ and $\mathscr{B}(\mathbf{x}^k) = \mathscr{B}(\widehat{\mathbf{x}})$. Thus, for $k \geq k_0$, all $\mathbf{x}^k$ that satisfy $\widehat{\mathbf{x}} \neq \mathbf{x}^{k-1}$ are generated by the conjugate gradient steps and, by the finite termination property of CG, there is $k \leq k_0 + n$ such that $\mathbf{x}^k = \widehat{\mathbf{x}}$. □

Unfortunately, the discretization of contact problems with a smooth contact interface typically results in the QP problems with a dual degenerate or nearly dual degenerate solution. The reason is that there can be the couples of points on the boundary of contact interface that are in contact but do not press each other. The solution of (8.1) which does not satisfy the strict complementarity condition is in Fig. 8.2.

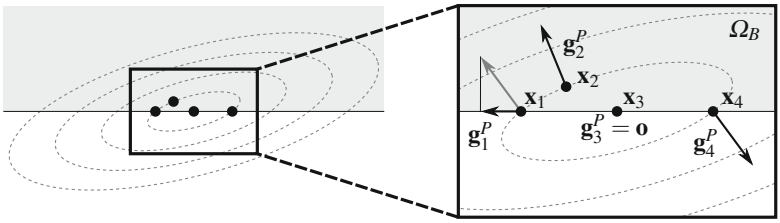

Fig. 8.2 Projected gradients near dual degenerate solution

A unique feature of MPRGP is that it preserves the finite termination property even in this case provided the balancing parameter Γ is sufficiently large. The result is a subject of the following theorem (see [1, Theorem 5.21]).

Theorem 8.2 *Let* $\{\mathbf{x}^k\}$ *denote the sequence generated by Algorithm 8.1 with*

$$\mathbf{x}^0 \in \Omega_B, \quad \Gamma \geq 3\left(\sqrt{\kappa(\mathsf{A})}+4\right), \quad \text{and} \quad \alpha \in (0, 2\|\mathsf{A}\|^{-1}]. \tag{8.24}$$

Then there is $k \geq 0$ *such that* $\mathbf{x}^k = \widehat{\mathbf{x}}$.

Let us recall that the finite termination property of the MPRGP algorithm with a dual degenerate solution and

$$\alpha \in (0, \|\mathsf{A}\|^{-1}]$$

has been proved for

$$\Gamma \geq 2\left(\sqrt{\kappa(\mathsf{A})}+1\right).$$

For the details see Dostál and Schöberl [7].

8.5 Implementation of MPRGP

In this section, we describe Algorithm 8.1 in the form which is convenient for implementation. To improve the efficiency of expansion steps, we include the feasible half-step introduced in Sect. 7.6.1, which is now associated with the expansion step.

Recall that it uses the trial conjugate gradient direction $\mathbf{p}^k$ which is generated before the expansion step is invoked. We propose to generate first

$$\mathbf{x}^{k+\frac{1}{2}} = \mathbf{x}^k - \alpha_f \mathbf{p}^k \quad \text{and} \quad \mathbf{g}^{k+\frac{1}{2}} = \mathbf{g}^k - \alpha_f \mathsf{A}\mathbf{p}^k,$$

where α_f denotes the feasible steplength for $\mathbf{p}^k$ defined by

$$\alpha_f = \min_{i=1,\dots,n} \{(x_i^k - \ell_i)/p_i^k,\ p_i^k > 0\},$$

and then define

$$\mathbf{x}^{k+1} = P_{\Omega_S}\left(\mathbf{x}^{k+\frac{1}{2}} - \alpha\boldsymbol{\varphi}(\mathbf{x}^{k+\frac{1}{2}})\right).$$

To preserve readability, we do not distinguish the generations of auxiliary vectors. The MPRGP algorithm with the feasible step reads as follows.

Algorithm 8.2 Modified proportioning with reduced gradient projections (MPRGP).

Given a symmetric positive definite matrix A *of the order n, n-vectors* $\mathbf{b}$, ℓ, $\Omega_B = \{\mathbf{x} :\ \mathbf{x} \geq \ell\}$, $\mathbf{x}^0 \in \Omega_B$.
Step 0. {*Initialization.*}
  Choose $\Gamma > 0$, $\alpha \in \left(0, 2\|\mathsf{A}\|^{-1}\right)$, *set* $k = 0$, $\mathbf{g} = \mathsf{A}\mathbf{x}^0 - \mathbf{b}$, $\mathbf{p} = \boldsymbol{\varphi}(\mathbf{x}^0)$
  while $\|\mathbf{g}^P(\mathbf{x}^k)\|$ *is not small*
    if $\|\boldsymbol{\beta}(\mathbf{x}^k)\|^2 \leq \Gamma^2 \widetilde{\boldsymbol{\varphi}}(\mathbf{x}^k)^T \boldsymbol{\varphi}(\mathbf{x}^k)$
Step 1. {*Proportional* $\mathbf{x}^k$. *Trial conjugate gradient step.*}
      $\alpha_{cg} = \mathbf{g}^T\mathbf{p}/\mathbf{p}^T\mathsf{A}\mathbf{p}$, $\mathbf{y} = \mathbf{x}^k - \alpha_{cg}\mathbf{p}$
      $\alpha_f = \max\{\alpha :\ \mathbf{x}^k - \alpha\mathbf{p} \in \Omega_B\} = \min\{(x_i^k - \ell_i)/p_i :\ p_i > 0\}$
      if $\alpha_{cg} \leq \alpha_f$
Step 2. {*Conjugate gradient step.*}
        $\mathbf{x}^{k+1} = \mathbf{y}$, $\mathbf{g} = \mathbf{g} - \alpha_{cg}\mathsf{A}\mathbf{p}$,
        $\beta = \boldsymbol{\varphi}(\mathbf{y})^T\mathsf{A}\mathbf{p}/\mathbf{p}^T\mathsf{A}\mathbf{p}$, $\mathbf{p} = \boldsymbol{\varphi}(\mathbf{y}) - \beta\mathbf{p}$
      else
Step 3. {*Expansion step.*}
        $\mathbf{x}^{k+\frac{1}{2}} = \mathbf{x}^k - \alpha_f\mathbf{p}$, $\mathbf{g} = \mathbf{g} - \alpha_f\mathsf{A}\mathbf{p}$
        $\mathbf{x}^{k+1} = P_{\Omega_B}(\mathbf{x}^{k+\frac{1}{2}} - \alpha\boldsymbol{\varphi}(\mathbf{x}^{k+\frac{1}{2}}))$
        $\mathbf{g} = \mathsf{A}\mathbf{x}^{k+1} - \mathbf{b}$, $\mathbf{p} = \boldsymbol{\varphi}(\mathbf{x}^{k+1})$
      end if
    else
Step 4. {*Proportioning step.*}
      $\mathbf{d} = \boldsymbol{\beta}(\mathbf{x}^k)$, $\alpha_{cg} = \mathbf{g}^T\mathbf{d}/\mathbf{d}^T\mathsf{A}\mathbf{d}$
      $\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_{cg}\mathbf{d}$, $\mathbf{g} = \mathbf{g} - \alpha_{cg}\mathsf{A}\mathbf{d}$, $\mathbf{p} = \boldsymbol{\varphi}(\mathbf{x}^{k+1})$
    end if
    $k = k + 1$
  end while
Step 5. {*Return (possibly inexact) solution.*}
  $\widetilde{\mathbf{x}} = \mathbf{x}^k$

For the choice of parameters, see Sect. 7.6.2.

8.6 Preconditioning

The performance of the CG-based methods can be improved by preconditioning described in Sect. 5.4. However, the application of preconditioning requires some care, as the *preconditioning transforms variables, turning the bound constraints into more general inequality constraints*. In this section, we present a basic strategy for the preconditioning of auxiliary linear problems.

Probably, the most straightforward preconditioning strategy which preserves the bound constraints is the preconditioning applied to the diagonal block $\mathsf{A}_{\mathscr{F}\mathscr{F}}$ of the Hessian matrix A in the conjugate gradient loop which minimizes the cost function f in the face defined by a free set $\mathscr{F}$. Such preconditioning requires that we are able to define for each diagonal block $\mathsf{A}_{\mathscr{F}\mathscr{F}}$ a regular matrix $\mathsf{M}(\mathscr{F})$ which satisfies the following two conditions. First, we require that $\mathsf{M}(\mathscr{F})$ approximates $\mathsf{A}_{\mathscr{F}\mathscr{F}}$ so that the convergence of the conjugate gradients method is significantly accelerated. The second condition requires that the solution of the system

$$\mathsf{M}(\mathscr{F})\mathbf{x} = \mathbf{y}$$

can be obtained easily. The preconditioners $\mathsf{M}(\mathscr{F})$ can be generated, e.g., by any of the methods described in Sect. 5.4.

Though the performance of the algorithm can be considerably improved by the preconditioning, the *preconditioning in face does not result in the improved bound on the rate of convergence*. The reason is that such preconditioning affects only the feasible conjugate gradient steps, leaving the expansion and proportioning steps without any preconditioning.

In probably the first application of preconditioning to the solution of bound-constrained problems [8], O'Leary considered two simple methods which can be used to obtain the preconditioner for $\mathsf{A}_{\mathscr{F}\mathscr{F}}$ from the preconditioner M which approximates A, namely,

$$\mathsf{M}(\mathscr{F}) = \mathsf{M}_{\mathscr{F}\mathscr{F}} \quad \text{and} \quad \mathsf{M}(\mathscr{F}) = \mathsf{L}_{\mathscr{F}\mathscr{F}}\mathsf{L}_{\mathscr{F}\mathscr{F}}^T,$$

where L denotes the factor of the Cholesky factorization $\mathsf{M} = \mathsf{L}\mathsf{L}^T$. It can be proved that whichever method of the preconditioning is used, the convergence bound for the conjugate gradient algorithm applied to the subproblems is at least as good as that of the conjugate gradient method applied to the original matrix [8].

To describe the MPRGP algorithm with the preconditioning in face, let us assume that we are given the preconditioner $\mathsf{M}(\mathscr{F})$ for each set of indices $\mathscr{F}$, and let us denote $\mathscr{F}_k = \mathscr{F}(\mathbf{x}^k)$ and $\mathscr{A}_k = \mathscr{A}(\mathbf{x}^k)$ for each vector $\mathbf{x}^k \in \Omega_B$. To simplify the description of the algorithm, let M_k denote the preconditioner corresponding to the face defined by $\mathscr{F}_k$ padded with zeros so that

$$\left[\mathsf{M}_k\right]_{\mathscr{F}\mathscr{F}} = \mathsf{M}(\mathscr{F}_k), \quad \left[\mathsf{M}_k\right]_{\mathscr{A}\mathscr{A}} = \mathsf{O}, \quad \left[\mathsf{M}_k\right]_{\mathscr{A}\mathscr{F}} = \left[\mathsf{M}_k\right]_{\mathscr{F}\mathscr{A}}^T = \mathsf{O},$$

and recall that $\mathsf{M}_k^\dagger$ denotes the Moore–Penrose inverse of M_k defined by

$$[\mathsf{M}_k^\dagger]_{\mathscr{F}\mathscr{F}} = \mathsf{M}(\mathscr{F}_k)^{-1}, \quad [\mathsf{M}_k^\dagger]_{\mathscr{A}\mathscr{A}} = \mathsf{O}, \quad [\mathsf{M}_k^\dagger]_{\mathscr{A}\mathscr{F}} = [\mathsf{M}_k^\dagger]^T_{\mathscr{F}\mathscr{A}} = \mathsf{O}.$$

In particular, it follows that

$$\mathsf{M}_k^\dagger \mathbf{g}(\mathbf{x}^k) = \mathsf{M}_k^\dagger \boldsymbol{\varphi}(\mathbf{x}^k).$$

The MPRGP algorithm with preconditioning in face reads as follows.

Algorithm 8.3 MPRGP with preconditioning in face.

Given a symmetric positive definite matrix A *of the order n, n-vectors* $\mathbf{b}$, ℓ, $\Omega_B = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq \ell\}$; *choose* $\mathbf{x}^0 \in \Omega_B$, $\Gamma > 0$, $\alpha \in (0, 2\|\mathsf{A}\|^{-1}]$, *and the rule which assigns to each* $\mathbf{x}^k \in \Omega_B$ *the preconditioner* M_k *which is SPD in the face defined by* $\mathscr{F}(\mathbf{x}^k)$.

Step 0. {*Initialization.*}

$\quad$ *Set* $k = 0$, $\mathbf{g} = \mathsf{A}\mathbf{x}^0 - \mathbf{b}$, $\mathbf{z} = \mathsf{M}_0^\dagger \mathbf{g}$, $\mathbf{p} = \mathbf{z}$

$\quad$ **while** $\|\mathbf{g}^P(\mathbf{x}^k)\|$ *is not small*

$\quad\quad$ **if** $\|\boldsymbol{\beta}(\mathbf{x}^k)\|^2 \leq \Gamma^2 \widetilde{\boldsymbol{\varphi}}(\mathbf{x}^k)^T \boldsymbol{\varphi}(\mathbf{x}^k)$

Step 1. {*Proportional* $\mathbf{x}^k$. *Trial conjugate gradient step.*}

$\quad\quad\quad$ $\alpha_{cg} = \mathbf{z}^T\mathbf{g}/\mathbf{p}^T\mathsf{A}\mathbf{p}$, $\mathbf{y} = \mathbf{x}^k - \alpha_{cg}\mathbf{p}$

$\quad\quad\quad$ $\alpha_f = \max\{\alpha : \mathbf{x}^k - \alpha\mathbf{p} \in \Omega_B\} = \min\{(x_i^k - \ell_i)/p_i : p_i > 0\}$

$\quad\quad\quad$ **if** $\alpha_{cg} \leq \alpha_f$

Step 2. {*Conjugate gradient step.*}

$\quad\quad\quad\quad$ $\mathbf{x}^{k+1} = \mathbf{y}$, $\mathbf{g} = \mathbf{g} - \alpha_{cg}\mathsf{A}\mathbf{p}$, $\mathbf{z} = \mathsf{M}_k^\dagger \mathbf{g}$

$\quad\quad\quad\quad$ $\beta = \mathbf{z}^T\mathsf{A}\mathbf{p}/\mathbf{p}^T\mathsf{A}\mathbf{p}$, $\mathbf{p} = \mathbf{z} - \beta\mathbf{p}$

$\quad\quad\quad$ **else**

Step 3. {*Expansion step.*}

$\quad\quad\quad\quad$ $\mathbf{x}^{k+\frac{1}{2}} = \mathbf{x}^k - \alpha_f\mathbf{p}$, $\mathbf{g} = \mathbf{g} - \alpha_f\mathsf{A}\mathbf{p}$

$\quad\quad\quad\quad$ $\mathbf{x}^{k+1} = P_{\Omega_B}\left(\mathbf{x}^{k+\frac{1}{2}} - \alpha\boldsymbol{\varphi}(\mathbf{x}^{k+\frac{1}{2}})\right)$

$\quad\quad\quad\quad$ $\mathbf{g} = \mathsf{A}\mathbf{x}^{k+1} - \mathbf{b}$, $\mathbf{z} = \mathsf{M}_{k+1}^\dagger \mathbf{g}$, $\mathbf{p} = \mathbf{z}$

$\quad\quad\quad$ **end** if

$\quad\quad$ **else**

Step 4. {*Proportioning step.*}

$\quad\quad\quad$ $\mathbf{d} = \boldsymbol{\beta}(\mathbf{x}^k)$, $\alpha_{cg} = \mathbf{g}^T\mathbf{d}/\mathbf{d}^T\mathsf{A}\mathbf{d}$

$\quad\quad\quad$ $\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_{cg}\mathbf{d}$, $\mathbf{g} = \mathbf{g} - \alpha_{cg}\mathsf{A}\mathbf{d}$, $\mathbf{z} = \mathsf{M}_{k+1}^\dagger \mathbf{g}$, $\mathbf{p} = \mathbf{z}$

$\quad\quad$ **end** if

$\quad\quad$ $k = k + 1$

$\quad$ **end** while

Step 5. {*Return (possibly inexact) solution.*}

$\quad$ $\widetilde{\mathbf{x}} = \mathbf{x}^k$

8.7 Comments and References

Since the conjugate gradient method was introduced in the celebrated paper by Hestenes and Stiefel [9] as a method for the solution of systems of linear equations, it seems that Polyak [10] was the first researcher who proposed to use the conjugate gradient method to minimize a quadratic cost function subject to bound constraints. Though Polyak assumed the auxiliary problems to be solved exactly, O'Leary [8] observed that this assumption can be replaced by refining the accuracy in the process of solution. In this way, she managed to reduce the number of iterations to about a half as compared with the algorithm using the exact solution. In spite of this, the convergence of these algorithms was supported by the arguments which gave only exponential bound on the number of matrix–vector multiplications that are necessary to reach the solution.

An important step forward was the development of algorithms with a rigorous convergence theory. On the basis of the results of Calamai and Moré [11], Moré and Toraldo [6] proposed an algorithm that also exploits the conjugate gradients and projections, but its convergence is driven by the gradient projections with the steplength satisfying the sufficient decrease condition (see, e.g., Nocedal and Wright [12]). The steplength is found, as in earlier algorithms, by possibly expensive backtracking. In spite of the iterative basis of their algorithm, the authors proved that their algorithm preserves the finite termination property of the original algorithm provided the solution satisfies the strict complementarity condition.

Friedlander, Martínez, Dostál, and their collaborators combined this result with an inexact solution of auxiliary problems [2, 3, 13, 14]. The concept of proportioning algorithm as presented here was introduced by Dostál in [3]. The convergence of the proportioning algorithm was driven by the proportioning step, leaving more room for the heuristic implementation of projections as compared with Moré and Toraldo [6]. The heuristics for implementation of the proportioning algorithm of Dostál [3] can be applied also to the MPRGP algorithm of Sect. 8.2. Comprehensive experiments and tests of heuristics can be found in Diniz-Ehrhardt, Gomes-Ruggiero, and Santos [15].

A common drawback of all above-mentioned strategies is possible backtracking in search of the gradient projection steplength and the lack of results on the rate of convergence. A key to further progress were the results by Schöberl [16, 17] and Dostál [18] on the decrease of the cost function along the projected-gradient path. (see also Sect. 6.6). It was observed by Dostál [19] that these results can be plugged into the proportioning algorithm in a way which preserves the rate of convergence. In our exposition of the MPRGP algorithm, we follow Dostál and Schöberl [7] and Dostál, Domorádová, and Sadowská [4]. See also the book [1].

The preconditioning in face was probably first considered by O'Leary [8]. The MPRGP with projector preconditioning is a key ingredient of the scalable algorithms for transient contact problems [20]. See also Chap. 13.

References

1. Dostál, Z.: Optimal Quadratic Programming Algorithms, with Applications to Variational Inequalities, 1st edn. Springer, New York (2009)
2. Friedlander, A., Martínez, J.M.: On the maximization of a concave quadratic function with box constraints. SIAM J. Optim. **4**, 177–192 (1994)
3. Dostál, Z.: Box constrained quadratic programming with proportioning and projections. SIAM J. Optim. **7**(3), 871–887 (1997)
4. Dostál, Z., Domorádová, M., Sadowská, M.: Superrelaxation in minimizing quadratic functions subject to bound constraints. Comput. Optim. Appl. **48**(1), 23–44 (2011)
5. van der Sluis, A., van der Vorst, H.A.: The rate of convergence of the conjugate gradients. Numer. Math. **48**, 543–560 (1986)
6. Moré, J.J., Toraldo, G.: On the solution of large quadratic programming problems with bound constraints. SIAM J. Optim. **1**, 93–113 (1991)
7. Dostál, Z., Schöberl, J.: Minimizing quadratic functions over non-negative cone with the rate of convergence and finite termination. Comput. Optim. Appl. **30**(1), 23–44 (2005)
8. O'Leary, D.P.: A generalised conjugate gradient algorithm for solving a class of quadratic programming problems. Linear Algebra Appl. **34**, 371–399 (1980)
9. Hestenes, M.R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Res. Natl. Bur. Stand. **49**, 409–436 (1952)
10. Polyak, B.T.: The conjugate gradient method in extremal problems. USSR Comput. Math. Math. Phys. **9**, 94–112 (1969)
11. Calamai, P.H., Moré, J.J.: Projected gradient methods for linearly constrained problems. Math. Program. **39**, 93–116 (1987)
12. Nocedal, J., Wright, S.F.: Numerical Optimization. Springer-Verlag, New York (2000)
13. Friedlander, A., Martínez, J.M., Raydan, M.: A new method for large scale box constrained quadratic minimization problems. Optim. Methods Softw. **5**, 57–74 (1995)
14. Bielschowski, R.H., Friedlander, A., Gomes, F.A.M., Martínez, J.M., Raydan, M.: An adaptive algorithm for bound constrained quadratic minimization. Invest. Oper. **7**, 67–102 (1997)
15. Diniz-Ehrhardt, M.A., Gomes-Ruggiero, M.A., Santos, S.A.: Numerical analysis of the leaving-face criterion in bound-constrained quadratic minimization. Optim. Methods Softw. **15**(1), 45–66 (2001)
16. Schöberl, J.: Solving the Signorini problem on the basis of domain decomposition techniques. Computing **60**(4), 323–344 (1998)
17. Schöberl, J.: Efficient contact solvers based on domain decomposition techniques. Comput. Math. Appl. **42**, 1217–1228 (2001)
18. Dostál, Z.: On the decrease of a quadratic function along the projected-gradient path. ETNA **31**, 25–59 (2008)
19. Dostál, Z.: A proportioning based algorithm for bound constrained quadratic programming with the rate of convergence. Numer. Algorithms **34**(2–4), 293–302 (2003)
20. Dostál, Z., Kozubek, T., Brzobohatý, T., Markopoulos, A., Vlach, O.: Scalable TFETI with optional preconditioning by conjugate projector for transient contact problems of elasticity. Comput. Methods Appl. Mech. Eng. **247–248**, 37–50 (2012)

Chapter 9
Solvers for Separable and Equality QP/QCQP Problems

We shall now use the results of our previous investigations to develop efficient algorithms for the minimization of strictly convex quadratic functions subject to possibly nonlinear *convex separable inequality constraints* and *linear equality constraints*

$$\min_{\mathbf{x}\in\Omega_{SE}} f(\mathbf{x}), \qquad f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathsf{A}\mathbf{x} - \mathbf{x}^T\mathbf{b}, \tag{9.1}$$

where

$$\Omega_{SE} = \{\mathbf{x} \in \mathbb{R}^n : \mathsf{B}\mathbf{x} = \mathbf{o} \text{ and } \mathbf{x} \in \Omega_S\}, \quad \Omega_S = \{\mathbf{x} \in \mathbb{R}^n : h_i(\mathbf{x}_i) \le 0,\ i = 1, \ldots, s\},$$

$\mathbf{b} \in \mathbb{R}^n$, h_i are convex functions, A is an $n \times n$ SPD matrix, and $\mathsf{B} \in \mathbb{R}^{m\times n}$. We are especially interested in the problems with bound and/or spherical and/or elliptic inequality constraints. We consider similar assumptions as in previous chapters. In particular, we assume $\Omega_{SE} \neq \emptyset$ and admit dependent rows of B. We assume that $\mathsf{B} \neq \mathsf{O}$ is not a full column rank matrix, so that $\operatorname{Ker}\mathsf{B} \neq \{\mathbf{o}\}$. Moreover, we assume that the constraints satisfy the Abadie constraint qualification introduced in Sect. 3.5.3, so that the solution can be characterized by the tangent cone, though we shall give some convergence results without this assumption. Observe that more general QP or QCQP programming problems can be reduced to (9.1) by duality, a suitable shift of variables, or by a modification of f.

The main idea of the algorithms that we develop here is to treat both of the sets of constraints separately. This approach enables us to use the ingredients of the algorithms developed in the previous chapters, such as the precision control of auxiliary problems. We restrict our attention to the SMALSE-M algorithm (Semi-Monotonic Augmented Lagrangian Algorithm for Separable and Equality constraints), which will be proved to have important optimality properties. We shall discuss separately the variant of SMALSE-M called SMALBE-M for the solution of QP problems with bound and equality constraints. The SMALSE-M and SMALBE-M algorithms are the key tools for the solution of contact problems with and without friction.

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3_9

9.1 KKT Conditions

Since Ω_{SE} is closed and convex and f is assumed to be strictly convex, the solution of problem (9.1) exists and is necessarily unique by Proposition 3.4. The conditions that are satisfied by the solution of (9.1) can be formulated by means of the augmented Lagrangian

$$L(\mathbf{x}, \boldsymbol{\lambda}_E, \boldsymbol{\mu}, \rho) = \frac{1}{2}\mathbf{x}^T\mathsf{A}\mathbf{x} - \mathbf{x}^T\mathbf{b} + \mathbf{x}^T\mathsf{B}^T\boldsymbol{\lambda}_E + \frac{\rho}{2}\|\mathsf{B}\mathbf{x}\|^2 + \sum_{i=1}^{s}\mu_i h_i(\mathbf{x}_i),$$

the gradient of which reads

$$\nabla_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}_E, \boldsymbol{\mu}, \rho) = (\mathsf{A} + \rho\mathsf{B}^T\mathsf{B})\mathbf{x} - \mathbf{b} + \mathsf{B}^T\boldsymbol{\lambda}_E + \sum_{i=1}^{s}\mu_i \nabla h_i(\mathbf{x}_i).$$

Using (3.47) and the Abadie constraint qualification, we get that a feasible vector $\mathbf{x} \in \Omega_{SE}$ is a solution of (9.1) if and only if there are $\boldsymbol{\lambda}_E \in \mathbb{R}^m$ and $\mu_1, \ldots, \mu_s$ such that

$$\nabla_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}_E, \boldsymbol{\mu}, \rho) = \mathbf{o}, \quad \mu_i \geq 0, \quad \text{and} \quad \mu_i h_i(\mathbf{x}_i) = 0, \quad i = 1, \ldots, s. \tag{9.2}$$

Having effective algorithms for the solution of QCQP problems with separable constraints, it is convenient to use explicitly the Lagrange multipliers only for the equality constraints, i.e., to use

$$L(\mathbf{x}, \boldsymbol{\lambda}_E, \rho) = \frac{1}{2}\mathbf{x}^T\mathsf{A}\mathbf{x} - \mathbf{x}^T\mathbf{b} + \mathbf{x}^T\mathsf{B}^T\boldsymbol{\lambda}_E + \frac{\rho}{2}\|\mathsf{B}\mathbf{x}\|^2.$$

Denoting by $\mathbf{g} = \mathbf{g}(\mathbf{x}, \boldsymbol{\lambda}_E, \rho)$ the gradient of the reduced augmented Lagrangian, so that

$$\mathbf{g} = \mathbf{g}(\mathbf{x}, \boldsymbol{\lambda}_E, \rho) = (\mathsf{A} + \rho\mathsf{B}^T\mathsf{B})\mathbf{x} - \mathbf{b} + \mathsf{B}^T\boldsymbol{\lambda}_E,$$

we get that $\mathbf{x} \in \Omega_{SE}$ is a solution of (9.1) if and only if there is $\boldsymbol{\lambda}_E \in \mathbb{R}^m$ such that

$$\mathbf{g}^P(\mathbf{x}, \boldsymbol{\lambda}_E, \rho) = \mathbf{o}, \tag{9.3}$$

where $\mathbf{g}^P$ is the projected gradient defined in Sect. 7.1 for the auxiliary problem

$$\min_{\mathbf{x} \in \Omega_S} L(\mathbf{x}, \boldsymbol{\lambda}_E, \rho), \quad \Omega_S = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{h}(\mathbf{x}) \leq \mathbf{o}\}. \tag{9.4}$$

However, condition (9.3) is sensitive to the curvature of the boundary, so we shall consider also an alternative condition (see Theorem 7.3)

$$\widetilde{\mathbf{g}}_{\alpha}^{P}(\mathbf{x}, \boldsymbol{\lambda}_E, \rho) = \mathbf{o} \tag{9.5}$$

with the reduced projected gradient $\widetilde{\mathbf{g}}_{\alpha}^{P}(\mathbf{x}, \boldsymbol{\lambda}_E, \rho)$ of Sect. 7.2 and $\alpha \in (0, 2\|\mathsf{A}\|^{-1})$.

9.2 Penalty and Method of Multipliers

Probably the most simple way how to exploit the algorithms of the previous chapters to the solution of (9.1) is to enhance the equality constraints into the objective function f by adding a suitable term which penalizes their violation. Thus the solution of (9.1) can be approximated by the solution of

$$\min_{\mathbf{x}\in\Omega_S} f_\rho(\mathbf{x}), \qquad f_\rho(\mathbf{x}) = f(\mathbf{x}) + \frac{\rho}{2}\|\mathsf{B}\mathbf{x}\|^2. \tag{9.6}$$

Intuitively, if the penalty parameter ρ is large, then the solution $\widehat{\mathbf{x}}_\rho$ of (9.6) can hardly be far from the solution of (9.1). Indeed, if ρ were infinite, then the minimizer of f_ρ would solve (9.1). Thus it is natural to expect that if ρ is sufficiently large, then the penalty approximation $\widehat{\mathbf{x}}_\rho$ is a suitable approximation to the solution $\widehat{\mathbf{x}}$ of (9.1). The effect of the penalty term is illustrated in Fig. 9.1. Notice that the penalty approximation is typically near the feasible set, but does not belong to it. That is why the penalty method is also called the *exterior penalty method*.

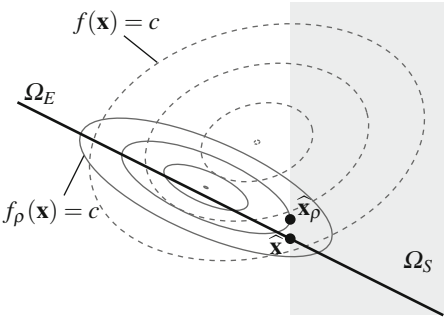

Fig. 9.1 The effect of the quadratic penalty

Because of its simplicity and intuitive appeal, the penalty method is often used in computations. However, a good approximation of the solution may require a very large penalty parameter, which can complicate computer implementation.

The remedy can be based on the observation that the solution $\widehat{\mathbf{x}}$ to (9.1) solves also

$$\min_{\mathbf{x}\in\Omega_S} L(\mathbf{x}, \boldsymbol{\lambda}_E, \rho) \tag{9.7}$$

with a suitable $\boldsymbol{\lambda}_E \in \mathbb{R}^m$. The point is that having a solution $\mathbf{x}_\rho$ of the penalized problem (9.7) with ρ and $\boldsymbol{\lambda}_E \in \mathbb{R}^m$, we can modify the linear term of L in such a

way that the minimum of the modified cost function *without the penalization term* with respect to $\mathbf{x} \in \Omega_S$ is achieved again at $\mathbf{x}_\rho$. The formula follows from

$$\min_{\mathbf{x}\in\Omega_S} L(\mathbf{x}, \boldsymbol{\lambda}_E, \rho) = \min_{\mathbf{x}\in\Omega_S} L(\mathbf{x}, \boldsymbol{\lambda}_E + \rho \mathsf{B}\mathbf{x}_\rho, 0).$$

Then we can find a better approximation by adding the penalization term to the modified cost function, and look for the minimizer of $L(\mathbf{x}, \boldsymbol{\lambda} + \rho\mathsf{B}\mathbf{x}_\rho, \rho)$. The result is the well-known classical *augmented Lagrangian algorithm*, also called the *method of multipliers*, which was proposed for general equality constraints by Hestenes [1] and Powell [2]. See also Bertsekas [3] or Glowinski and Le Tallec [4].

9.3 SMALSE-M

The following algorithm is a modification of the algorithm proposed by Conn, Gould, and Toint [5] for the minimization of more general cost functions subject to bound and equality constraints in their LANCELOT package. The SMALSE-M algorithm presented here differs from that used in LANCELOT by the adaptive precision control introduced by Hager [6] and Dostál, Friedlander, and Santos [7], by using a fixed regularization parameter, and by the control of the parameter M_k. Since we use only the multipliers for the equality constraints, we denote them by $\boldsymbol{\lambda}$, i.e., $\boldsymbol{\lambda} = \boldsymbol{\lambda}_E$. The complete SMALSE-M algorithm reads as follows.

Algorithm 9.1 Semimonotonic augmented Lagrangians for separable and equality constrained QCQP problems (SMALSE-M).

Given an SPD matrix $\mathsf{A} \in \mathbb{R}^{n\times n}$, $\mathsf{B} \in \mathbb{R}^{m\times n}$, *n-vector* $\mathbf{b}$, *constraints* $\mathbf{h}$.
Step 0. {*Initialization.*}
Choose $\eta > 0,\ 0 < \beta < 1,\ M_{-1} = 0,\ M_0 > 0,\ \rho > 0,\ \boldsymbol{\lambda}^0 \in \mathbb{R}^m$
for $k = 0, 1, 2, \ldots$
Step 1. {*Inner iteration with adaptive precision control.*}
Find $\mathbf{x}^k \in \Omega_S$ *such that*

$$\|\mathbf{g}^P(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho)\| \le \min\{M_k \|\mathsf{B}\mathbf{x}^k\|, \eta\} \tag{9.8}$$

Step 2. {*Updating the Lagrange multipliers.*}

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \rho\mathsf{B}\mathbf{x}^k \tag{9.9}$$

Step 3. {*Update M provided the increase of the Lagrangian is not sufficient.*}
if $M_k = M_{k-1}$ *and*

$$L(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho) < L(\mathbf{x}^{k-1}, \boldsymbol{\lambda}^{k-1}, \rho) + \frac{\rho}{2}\|\mathsf{B}\mathbf{x}^k\|^2 \tag{9.10}$$

$M_{k+1} = \beta M_k$
else
$M_{k+1} = M_k$
end else if
end for

In Step 1 we can use any algorithm for minimizing strictly convex quadratic functions subject to separable constraints provided it guarantees the convergence of projected gradients to zero. Our optimality theory requires that the algorithm in the inner loop has the rate of convergence in terms of bounds on the spectrum of A, such as the MPGP Algorithm 7.2 or MPRGP Algorithm 8.2. A *stopping criterion* should either follow Step 1 or be enhanced in Step 1. A natural choice is the *relative precision*

$$\|\mathbf{g}^P\| \le \varepsilon_g \|\mathbf{b}\| \quad \text{and} \quad \varepsilon_e \|\mathbf{b}\| \tag{9.11}$$

prescribed by small parameters $\varepsilon_g > 0$ and $\varepsilon_e > 0$.

The SMALSE algorithm requires four parameters M_0, β, η, ρ. The algorithm is not very sensitive to the choice of β and η and adjusts properly the balancing parameter M. A larger value of ρ increases the rate of convergence of SMALSE at the cost of slowing down the rate of convergence of the algorithm in the inner loop. Some hints concerning the initialization of the parameters and the implementation of SMALSE-M can be found in Sect. 9.14.

The next lemma shows that Algorithm 9.1 is well defined, i.e., any algorithm for the solution of auxiliary problems in Step 1 that guarantees the convergence of projected gradients to zero generates either $\mathbf{x}^k$ which satisfies (9.5) in a finite number of steps or the iterates which converge to the solution of (9.1).

Lemma 9.1 *Let $M > 0$, $\boldsymbol{\lambda} \in \mathbb{R}^m$, $\eta > 0$, and $\rho \ge 0$ be given. Let $\{\mathbf{y}^k\} \in \Omega_S$ denote any sequence such that*

$$\widehat{\mathbf{y}} = \lim_{k\to\infty} \mathbf{y}^k = \arg\min_{\mathbf{y}\in\Omega_S} L(\mathbf{y}, \boldsymbol{\lambda}, \rho)$$

and let $\mathbf{g}^P(\mathbf{y}^k, \boldsymbol{\lambda}, \rho)$ converges to zero vector. Then $\{\mathbf{y}^k\}$ either converges to the unique solution $\widehat{\mathbf{x}}$ of problem (9.1), or there is an index k such that

$$\|\mathbf{g}^P(\mathbf{y}^k, \boldsymbol{\lambda}, \rho)\| \le \min\{M\|\mathsf{B}\mathbf{y}^k\|, \eta\}. \tag{9.12}$$

Proof If (9.12) does not hold for any k, then

$$\|\mathbf{g}^P(\mathbf{y}^k, \boldsymbol{\lambda}, \rho)\| > M\|\mathsf{B}\mathbf{y}^k\|$$

for any k. Since $\mathbf{g}^P(\mathbf{y}^k, \boldsymbol{\lambda}, \rho)$ converges to zero vector by the assumption, it follows that $\|\mathsf{B}\mathbf{y}^k\|$ converges to zero. Thus $\mathsf{B}\widehat{\mathbf{y}} = \mathbf{o}$ and

$$\mathbf{g}^P(\widehat{\mathbf{y}}, \boldsymbol{\lambda}, \rho) = \mathbf{o}.$$

It follows that $\widehat{\mathbf{y}}$ satisfies the KKT conditions (9.2) and $\widehat{\mathbf{y}} = \widehat{\mathbf{x}}$. □

Remark 9.1 In Step 3, we can replace the update rule to

$$M_{k+1} = \beta M_k \ \text{ or } \ \rho = \rho/\beta.$$

We do not elaborate this option here.

9.4 Inequalities Involving the Augmented Lagrangian

In this section we establish basic inequalities which relate the bound on the norm of the projected gradient $\mathbf{g}^P$ of the augmented Lagrangian L to the values of L. These inequalities will be the key ingredients in the proof of convergence and other analysis concerning Algorithm 9.1.

Lemma 9.2 *Let* $\mathbf{x}, \mathbf{y} \in \Omega_S$, $\boldsymbol{\lambda} \in \mathbb{R}^m$, $\rho > 0$, $\eta > 0$, *and* $M > 0$. *Let* $\lambda_{\min}$ *denote the least eigenvalue of* A *and* $\widetilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda} + \rho \mathsf{B}\mathbf{x}$.
(i) If

$$\|\mathbf{g}^P(\mathbf{x}, \boldsymbol{\lambda}, \rho)\| \le M \|\mathsf{B}\mathbf{x}\|, \tag{9.13}$$

then

$$L(\mathbf{y}, \widetilde{\boldsymbol{\lambda}}, \rho) \ge L(\mathbf{x}, \boldsymbol{\lambda}, \rho) + \frac{1}{2}\left(\rho - \frac{M^2}{\lambda_{\min}}\right)\|\mathsf{B}\mathbf{x}\|^2 + \frac{\rho}{2}\|\mathsf{B}\mathbf{y}\|^2. \tag{9.14}$$

(ii) If

$$\|\mathbf{g}^P(\mathbf{x}, \boldsymbol{\lambda}, \rho)\| \le \eta, \tag{9.15}$$

then

$$L(\mathbf{y}, \widetilde{\boldsymbol{\lambda}}, \rho) \ge L(\mathbf{x}, \boldsymbol{\lambda}, \rho) + \frac{\rho}{2}\|\mathsf{B}\mathbf{x}\|^2 + \frac{\rho}{2}\|\mathsf{B}\mathbf{y}\|^2 - \frac{\eta^2}{2\lambda_{\min}}. \tag{9.16}$$

(iii) If $\mathbf{z}_0 \in \Omega_{SE}$ *and* (9.15), *then*

$$L(\mathbf{x}, \boldsymbol{\lambda}, \rho) \le f(\mathbf{z}_0) + \frac{\eta^2}{2\lambda_{\min}}. \tag{9.17}$$

Proof Let us denote $\boldsymbol{\delta} = \mathbf{y} - \mathbf{x}$ and $\mathsf{A}_\rho = \mathsf{A} + \rho \mathsf{B}^T\mathsf{B}$ and recall that by the assumptions and Lemma 7.1

$$\mathbf{g}^T(\mathbf{y} - \mathbf{x}) \ge (\mathbf{g}^P)^T(\mathbf{y} - \mathbf{x}).$$

Using

$$L(\mathbf{x}, \widetilde{\boldsymbol{\lambda}}, \rho) = L(\mathbf{x}, \boldsymbol{\lambda}, \rho) + \rho\|\mathsf{B}\mathbf{x}\|^2 \ \text{ and } \ \mathbf{g}(\mathbf{x}, \widetilde{\boldsymbol{\lambda}}, \rho) = \mathbf{g}(\mathbf{x}, \boldsymbol{\lambda}, \rho) + \rho\mathsf{B}^T\mathsf{B}\mathbf{x},$$

we get

$$\begin{aligned} L(\mathbf{y}, \widetilde{\boldsymbol{\lambda}}, \rho) &= L(\mathbf{x}, \widetilde{\boldsymbol{\lambda}}, \rho) + \boldsymbol{\delta}^T \mathbf{g}(\mathbf{x}, \widetilde{\boldsymbol{\lambda}}, \rho) + \frac{1}{2}\boldsymbol{\delta}^T \mathsf{A}_\rho \boldsymbol{\delta} \\ &= L(\mathbf{x}, \boldsymbol{\lambda}, \rho) + \boldsymbol{\delta}^T \mathbf{g}(\mathbf{x}, \boldsymbol{\lambda}, \rho) + \frac{1}{2}\boldsymbol{\delta}^T \mathsf{A}_\rho \boldsymbol{\delta} + \rho \boldsymbol{\delta}^T \mathsf{B}^T \mathsf{B}\mathbf{x} + \rho \|\mathsf{B}\mathbf{x}\|^2 \\ &\geq L(\mathbf{x}, \boldsymbol{\lambda}, \rho) + \boldsymbol{\delta}^T \mathbf{g}^P(\mathbf{x}, \boldsymbol{\lambda}, \rho) + \frac{1}{2}\boldsymbol{\delta}^T \mathsf{A}_\rho \boldsymbol{\delta} + \rho \boldsymbol{\delta}^T \mathsf{B}^T \mathsf{B}\mathbf{x} + \rho \|\mathsf{B}\mathbf{x}\|^2 \\ &\geq L(\mathbf{x}, \boldsymbol{\lambda}, \rho) + \boldsymbol{\delta}^T \mathbf{g}^P(\mathbf{x}, \boldsymbol{\lambda}, \rho) + \frac{\lambda_{\min}}{2}\|\boldsymbol{\delta}\|^2 + \frac{\rho}{2}\|\mathsf{B}\boldsymbol{\delta}\|^2 + \rho \boldsymbol{\delta}^T \mathsf{B}^T \mathsf{B}\mathbf{x} \\ &\quad + \rho \|\mathsf{B}\mathbf{x}\|^2. \end{aligned}$$

Noticing that

$$\frac{\rho}{2}\|\mathsf{B}\mathbf{y}\|^2 = \frac{\rho}{2}\|\mathsf{B}(\boldsymbol{\delta} + \mathbf{x})\|^2 = \rho \boldsymbol{\delta}^T \mathsf{B}^T \mathsf{B}\mathbf{x} + \frac{\rho}{2}\|\mathsf{B}\boldsymbol{\delta}\|^2 + \frac{\rho}{2}\|\mathsf{B}\mathbf{x}\|^2,$$

we get

$$\begin{aligned} L(\mathbf{y}, \widetilde{\boldsymbol{\lambda}}, \rho) \geq & L(\mathbf{x}, \boldsymbol{\lambda}, \rho) + \boldsymbol{\delta}^T \mathbf{g}^P(\mathbf{x}, \boldsymbol{\lambda}, \rho) \\ & + \frac{\lambda_{\min}}{2}\|\boldsymbol{\delta}\|^2 + \frac{\rho}{2}\|\mathsf{B}\mathbf{x}\|^2 + \frac{\rho}{2}\|\mathsf{B}\mathbf{y}\|^2. \end{aligned} \tag{9.18}$$

Using (9.13) and simple manipulations then yields

$$\begin{aligned} L(\mathbf{y}, \widetilde{\boldsymbol{\lambda}}, \rho) &\geq L(\mathbf{x}, \boldsymbol{\lambda}, \rho) - M\|\boldsymbol{\delta}\|\|\mathsf{B}\mathbf{x}\| + \frac{\lambda_{\min}}{2}\|\boldsymbol{\delta}\|^2 + \frac{\rho}{2}\|\mathsf{B}\mathbf{x}\|^2 + \frac{\rho}{2}\|\mathsf{B}\mathbf{y}\|^2 \\ &= L(\mathbf{x}, \boldsymbol{\lambda}, \rho) + \left(\frac{\lambda_{\min}}{2}\|\boldsymbol{\delta}\|^2 - M\|\boldsymbol{\delta}\|\|\mathsf{B}\mathbf{x}\| + \frac{M^2\|\mathsf{B}\mathbf{x}\|^2}{2\lambda_{\min}}\right) \\ &\quad - \frac{M^2\|\mathsf{B}\mathbf{x}\|^2}{2\lambda_{\min}} + \frac{\rho}{2}\|\mathsf{B}\mathbf{x}\|^2 + \frac{\rho}{2}\|\mathsf{B}\mathbf{y}\|^2 \\ &\geq L(\mathbf{x}, \boldsymbol{\lambda}, \rho) + \frac{1}{2}\left(\rho - \frac{M^2}{\lambda_{\min}}\right)\|\mathsf{B}\mathbf{x}\|^2 + \frac{\rho}{2}\|\mathsf{B}\mathbf{y}\|^2. \end{aligned}$$

This proves (i).
(ii) If we assume that (9.15) holds, then by (9.18)

$$\begin{aligned} L(\mathbf{y}, \widetilde{\boldsymbol{\lambda}}, \rho) &\geq L(\mathbf{x}, \boldsymbol{\lambda}, \rho) - \|\boldsymbol{\delta}\|\eta + \frac{\lambda_{\min}}{2}\|\boldsymbol{\delta}\|^2 + \frac{\rho}{2}\|\mathsf{B}\mathbf{x}\|^2 + \frac{\rho}{2}\|\mathsf{B}\mathbf{y}\|^2 \\ &\geq L(\mathbf{x}, \boldsymbol{\lambda}, \rho) + \frac{\rho}{2}\|\mathsf{B}\mathbf{x}\|^2 + \frac{\rho}{2}\|\mathsf{B}\mathbf{y}\|^2 - \frac{\eta^2}{2\lambda_{\min}}. \end{aligned}$$

This proves (ii).

(iii) Let $\widehat{\mathbf{z}}$ denote the solution of the auxiliary problem

$$\text{minimize} \;\; L(\mathbf{z}, \boldsymbol{\lambda}, \rho) \;\; \text{s.t.} \;\; \mathbf{z} \geq \ell, \tag{9.19}$$

let $\mathbf{z}_0 \in \Omega_{SE}$ so that $\mathsf{B}\mathbf{z}_0 = \mathbf{o}$, and let $\widehat{\boldsymbol{\delta}} = \widehat{\mathbf{z}} - \mathbf{x}$. If (9.15) holds, then

$$\begin{aligned} 0 \geq L(\widehat{\mathbf{z}}, \boldsymbol{\lambda}, \rho) - L(\mathbf{x}, \boldsymbol{\lambda}, \rho) &= \widehat{\boldsymbol{\delta}}^T \mathbf{g}(\mathbf{x}, \boldsymbol{\lambda}, \rho) + \frac{1}{2}\widehat{\boldsymbol{\delta}}^T \mathsf{A}_\rho \widehat{\boldsymbol{\delta}} \\ &\geq \widehat{\boldsymbol{\delta}}^T \mathbf{g}^P(\mathbf{x}, \boldsymbol{\lambda}, \rho) + \frac{1}{2}\widehat{\boldsymbol{\delta}}^T \mathsf{A}_\rho \widehat{\boldsymbol{\delta}} \geq -\|\widehat{\boldsymbol{\delta}}\|\eta + \frac{1}{2}\lambda_{\min}\|\widehat{\boldsymbol{\delta}}\|^2 \geq -\frac{\eta^2}{2\lambda_{\min}}. \end{aligned}$$

Since $L(\widehat{\mathbf{z}}, \boldsymbol{\lambda}, \rho) \leq L(\mathbf{z}_0, \boldsymbol{\lambda}, \rho) = f(\mathbf{z}_0)$, we conclude that

$$L(\mathbf{x}, \boldsymbol{\lambda}, \rho) \leq L(\mathbf{x}, \boldsymbol{\lambda}, \rho) - L(\widehat{\mathbf{z}}, \boldsymbol{\lambda}, \rho) + f(\mathbf{z}_0) \leq f(\mathbf{z}_0) + \frac{\eta^2}{2\lambda_{\min}}. \qquad \square$$

9.5 Monotonicity and Feasibility

Now we shall translate the results on the relations that are satisfied by the augmented Lagrangian into the relations concerning the iterates generated by SMALSE-M.

Lemma 9.3 *Let $\{\mathbf{x}^k\}$, $\{\boldsymbol{\lambda}^k\}$, and M_k be generated by Algorithm 9.1 for solving (9.1) with $\eta > 0$, $0 < \beta < 1$, $M_0 > 0$, ρ, and $\boldsymbol{\lambda}^0 \in \mathbb{R}^m$. Let $\lambda_{\min}$ denote the least eigenvalue of the Hessian A of the quadratic function f.*

(i) If $k > 0$, $M_k = M_{k+1}$, and

$$M_k \leq \sqrt{\lambda_{\min}\rho}, \tag{9.20}$$

then

$$L(\mathbf{x}^{k+1}, \boldsymbol{\lambda}^{k+1}, \rho) \geq L(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho) + \frac{\rho}{2}\|\mathsf{B}\mathbf{x}^{k+1}\|^2. \tag{9.21}$$

(ii) For any $k \geq 0$

$$\begin{aligned} L(\mathbf{x}^{k+1}, \boldsymbol{\lambda}^{k+1}, \rho) \geq L(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho) &+ \frac{\rho}{2}\|\mathsf{B}\mathbf{x}^k\|^2 \\ &+ \frac{\rho}{2}\|\mathsf{B}\mathbf{x}^{k+1}\|^2 - \frac{\eta^2}{2\lambda_{\min}}. \end{aligned} \tag{9.22}$$

(iii) For any $k \geq 0$ and $\mathbf{z}_0 \in \Omega_{SE}$

$$L(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho) \leq f(\mathbf{z}_0) + \frac{\eta^2}{2\lambda_{\min}}. \tag{9.23}$$

Proof In Lemma 9.2, let us substitute $\mathbf{x} = \mathbf{x}^k$, $\boldsymbol{\lambda} = \boldsymbol{\lambda}^k$, $M = M_k$, and $\mathbf{y} = \mathbf{x}^{k+1}$, so that inequalities (9.13) and (9.15) hold by (9.8) and $\widetilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda}^{k+1}$. Thus we get Lemma 9.3 from Lemma 9.2.

If (9.20) holds, we can use (9.14) to get (9.21). Similarly, inequalities (9.22) and (9.23) can be obtained by the substitution into Lemma 9.2 (ii) and (iii), respectively. □

Theorem 9.1 *Let* $\{\mathbf{x}^k\}$, $\{\boldsymbol{\lambda}^k\}$, *and* $\{M_k\}$ *be generated by Algorithm 9.1 for the solution of (9.1) with* $\eta > 0$, $0 < \beta < 1$, $M_0 > 0$, $\rho > 0$, *and* $\boldsymbol{\lambda}^0 \in \mathbb{R}^m$. *Let* $\lambda_{\min}$ *denote the least eigenvalue of the Hessian* A *of the cost function* f, *and let* $p \geq 0$ *denote the smallest integer such that* $\rho \geq (\beta^p M_0)^2/\lambda_{\min}$. *Then the following statements hold.*

(i) There is k_0 *such that*

$$\min\{M_0, \beta\sqrt{\rho\lambda_{\min}}\} \leq M_{k_0} = M_{k_0+1} = M_{k_0+2} = \cdots \tag{9.24}$$

(ii) If $\mathbf{z}_0 \in \Omega_{SE}$, *then*

$$\frac{\rho}{2}\sum_{k=1}^{\infty} \|\mathsf{B}\mathbf{x}^k\|^2 \leq f(\mathbf{z}_0) - L(\mathbf{x}^0, \boldsymbol{\lambda}^0, \rho) + (1+p)\frac{\eta^2}{2\lambda_{\min}}. \tag{9.25}$$

(iii)

$$\lim_{k\to 0} \mathbf{g}^P(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho) = \mathbf{o} \quad \text{and} \quad \lim_{k\to 0} \mathsf{B}\mathbf{x}^k = \mathbf{o}. \tag{9.26}$$

Proof Let $p \geq 0$ denote the smallest integer such that $\rho \geq (\beta^p M_0)^2/\lambda_{\min}$ and let $\mathscr{I} \subseteq \{1, 2, \dots\}$ denote a possibly empty set of the indices k_i such that $M_{k_i} < M_{k_i-1}$. Using Lemma 9.3(i), $M_{k_i} = \beta M_{k_i-1} = \beta^i M_0$ for $k_i \in \mathscr{I}$, and $\rho \geq (\beta^p M_0)^2/\lambda_{\min}$, we conclude that there is no k such that $M_k < \beta^p M_0$. Thus $\mathscr{I}$ has at most p elements and (9.24) holds. By the definition of Step 3, if $k > 0$, then either $k \notin \mathscr{I}$ and

$$\frac{\rho}{2}\|\mathsf{B}\mathbf{x}^k\|^2 \leq L(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho) - L(\mathbf{x}^{k-1}, \boldsymbol{\lambda}^{k-1}, \rho),$$

or $k \in \mathscr{I}$ and by (9.22)

$$\begin{aligned}\frac{\rho}{2}\|\mathsf{B}\mathbf{x}^k\|^2 &\leq \frac{\rho}{2}\|\mathsf{B}\mathbf{x}^{k-1}\|^2 + \frac{\rho}{2}\|\mathsf{B}\mathbf{x}^k\|^2 \\ &\leq L(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho) - L(\mathbf{x}^{k-1}, \boldsymbol{\lambda}^{k-1}, \rho) + \frac{\eta^2}{2\lambda_{\min}}.\end{aligned}$$

Summing up the appropriate cases of the last two inequalities for $k = 1, \dots, j$ and taking into account that $\mathscr{I}$ has at most p elements, we get

$$\sum_{k=1}^{j} \frac{\rho}{2}\|\mathsf{B}\mathbf{x}^k\|^2 \leq L(\mathbf{x}^j, \boldsymbol{\lambda}^j, \rho) - L(\mathbf{x}^0, \boldsymbol{\lambda}^0, \rho) + p\frac{\eta^2}{2\lambda_{\min}}. \tag{9.27}$$

To get (9.25), it is enough to replace $L(\mathbf{x}^j, \boldsymbol{\lambda}^j, \rho)$ by the upper bound (9.23). The relations (9.26) are easy corollaries of (9.25) and the definition of Step 1. □

9.6 Boundedness

The first step toward the proof of convergence of SMALSE-M is to show that $\mathbf{x}^k$ are bounded.

Proposition 9.1 *Let $\{\mathbf{x}^k\}$ and $\{\boldsymbol{\lambda}^k\}$ be generated by Algorithm 9.1 for the solution of (9.1) with $\eta > 0$, $0 < \beta < 1$, $M_0 > 0$, $\rho > 0$, and $\boldsymbol{\lambda}^0 \in \mathbb{R}^m$. For each $i \in \{1, \ldots, s\}$, let $\mathscr{I}_i$ denote the indices of the components of $\mathbf{x}$ associated with the argument of the constraint function h_i, so that $\mathbf{x}_i = [\mathbf{x}]_{\mathscr{J}_i}$, and let us define*

$$\mathscr{A}(\mathbf{x}) = \cup_{i \in \mathscr{A}(\mathbf{x})} \mathscr{I}_i, \quad \widetilde{\mathscr{F}}(\mathbf{x}) = \mathscr{N} \setminus \widetilde{\mathscr{A}}(\mathbf{x}), \quad \mathbf{x}_{\mathscr{F}} = \mathbf{x}_{\widetilde{\mathscr{F}}(\mathbf{x})}, \quad \mathbf{x}_{\mathscr{A}} = \mathbf{x}_{\widetilde{\mathscr{A}}(\mathbf{x})}.$$

Let the boundary of Ω_S is bounded, i.e., there is $C > 0$ such that for any $\mathbf{x} \in \Omega_S$

$$\|\mathbf{x}_{\mathscr{A}(\mathbf{x})}\|^2 \le C.$$

Then $\{\mathbf{x}^k\}$ is bounded.

Proof Since there is only a finite number of different subsets $\mathscr{F}$ of the set of all indices $\mathscr{N} = \{1, \ldots, n\}$ and $\{\mathbf{x}^k\}$ is bounded if and only if $\{\mathbf{x}^k_{\mathscr{F}}\}$ is bounded for any $\mathscr{F} \subseteq \mathscr{N}$, we can restrict our attention to the analysis of the infinite subsequences $\{\mathbf{x}^k_{\mathscr{F}} : \widetilde{\mathscr{F}}(\mathbf{x}^k) = \mathscr{F}\}$ that are defined by the nonempty subsets $\mathscr{F}$ of $\mathscr{N}$.

Let $\mathscr{F} \subseteq \mathscr{N}$, $\mathscr{F} \neq \emptyset$, let $\mathscr{K} = \{k : \widetilde{\mathscr{F}}(\mathbf{x}^k) = \mathscr{F}\}$ be infinite, and denote

$$\mathscr{A} = \mathscr{N} \setminus \mathscr{F}, \quad \mathsf{H} = \mathsf{A} + \rho_k \mathsf{B}^T \mathsf{B}.$$

We get

$$\mathbf{g}^k = \mathbf{g}(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho) = \mathsf{H}\mathbf{x}^k + \mathsf{B}^T \boldsymbol{\lambda}^k - \mathbf{b}$$

and

$$\begin{bmatrix} \mathsf{H}_{\mathscr{F}\mathscr{F}} & \mathsf{B}^T_{*\mathscr{F}} \\ \mathsf{B}_{*\mathscr{F}} & \mathsf{O} \end{bmatrix} \begin{bmatrix} \mathbf{x}^k_{\mathscr{F}} \\ \boldsymbol{\lambda}^k \end{bmatrix} = \begin{bmatrix} \mathbf{g}^k_{\mathscr{F}} + \mathbf{b}_{\mathscr{F}} - \mathsf{H}_{\mathscr{F}A}\mathbf{x}^k_{\mathscr{A}} \\ \mathsf{B}_{*\mathscr{F}}\mathbf{x}^k_{\mathscr{F}} \end{bmatrix}. \tag{9.28}$$

Since for $k \in \mathscr{K}$

$$\mathsf{B}_{*\mathscr{F}}\mathbf{x}^k_{\mathscr{F}} = \mathsf{B}\mathbf{x}^k - \mathsf{B}_{*\mathscr{A}}\mathbf{x}_{\mathscr{A}}, \quad \|\mathbf{g}^k_{\mathscr{F}}\| = \|\mathbf{g}_{\mathscr{F}}(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho_k)\| \le \|\mathbf{g}^P(\mathbf{x}^k\boldsymbol{\lambda}^k, \rho_k)\|,$$

and both $\|\mathbf{g}^P(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho)\|$ and $\|\mathsf{B}\mathbf{x}^k\|$ converge to zero by the definition of $\mathbf{x}^k$ in Step 1 of Algorithm 9.1 and (9.25), the right-hand side of (9.28) is bounded. Since $\mathsf{H}_{\mathscr{F}F}$ is nonsingular, it is easy to check that the matrix of the system (9.28) is nonsingular

when $\mathsf{B}_{*\mathscr{F}}$ is a full row rank matrix. It simply follows that both $\{\mathbf{x}^k\}$ and $\{\boldsymbol{\lambda}^k\}$ are bounded provided the matrix of (9.28) is nonsingular.

If $\mathsf{B}_{*\mathscr{F}}$ is not a full row rank matrix, then its rank r satisfies $r < m$, and by the RSVD formula (2.28) there are matrices

$$\mathsf{U} \in \mathbb{R}^{m\times r}, \quad \mathsf{V} \in \mathbb{R}^{n\times r}, \quad \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r), \quad \sigma_i > 0, \quad \mathsf{U}^T\mathsf{U} = \mathsf{I}, \quad \mathsf{V}^T\mathsf{V} = \mathsf{I},$$

such that $\mathsf{B}_{*\mathscr{F}} = \mathsf{U}\Sigma\mathsf{V}^T$. Thus we can define the full row rank matrix $\hat{\mathsf{B}}_{*\mathscr{F}} = \Sigma\mathsf{V}^T$ that satisfies

$$\|\hat{\mathsf{B}}_{*\mathscr{F}}\mathbf{x}_{\mathscr{F}}\| = \|\Sigma\mathsf{V}^T\mathbf{x}_{\mathscr{F}}\| = \|\mathsf{U}\Sigma\mathsf{V}^T\mathbf{x}_{\mathscr{F}}\| = \|\mathsf{B}_{*\mathscr{F}}\mathbf{x}_{\mathscr{F}}\|$$

for any vector $\mathbf{x}$. Let us assign to any $\boldsymbol{\lambda} \in \mathbb{R}^m$ the vector $\hat{\boldsymbol{\lambda}} = \mathsf{U}^T\boldsymbol{\lambda}$, so that

$$\hat{\mathsf{B}}^T_{*\mathscr{F}}\hat{\boldsymbol{\lambda}} = \mathsf{V}\Sigma\mathsf{U}^T\boldsymbol{\lambda} = \mathsf{B}^T_{*\mathscr{F}}\boldsymbol{\lambda}.$$

Using the latter identity and (9.28), we get the system

$$\begin{bmatrix} \mathsf{H}_{\mathscr{F}F} & \hat{\mathsf{B}}^T_{*\mathscr{F}} \\ \hat{\mathsf{B}}_{*\mathscr{F}} & \mathsf{O} \end{bmatrix} \begin{bmatrix} \mathbf{x}^k_{\mathscr{F}} \\ \hat{\boldsymbol{\lambda}}^k \end{bmatrix} = \begin{bmatrix} \mathbf{g}^k_{\mathscr{F}} + \mathbf{b}_{\mathscr{F}} - \mathsf{H}_{\mathscr{F}A}\mathbf{x}_{\mathscr{A}} \\ \hat{\mathsf{B}}_{*\mathscr{F}}\mathbf{x}^k_{\mathscr{F}} \end{bmatrix} \tag{9.29}$$

with a nonsingular matrix. The right-hand side of (9.29) being bounded due to $\|\hat{\mathsf{B}}_{*\mathscr{F}}\mathbf{x}^k_{\mathscr{F}}\| = \|\mathsf{B}_{*\mathscr{F}}\mathbf{x}^k_{\mathscr{F}}\|$, we conclude that the set $\{\mathbf{x}^k_{\mathscr{F}} : \mathscr{F}(\mathbf{x}^k) = \mathscr{F}\}$ is bounded. See also Dostál and Kučera [8] or Dostál and Kozubek [9]. □

9.7 Convergence

Now we are ready to prove the main convergence results of this chapter. To describe them effectively, let $\mathscr{F} = \widetilde{\mathscr{F}}(\widehat{\mathbf{x}})$ denote the set of indices of the variables that are involved in the free set of a unique solution $\widehat{\mathbf{x}}$ (see Proposition 9.1 for formal definition of $\widetilde{\mathscr{F}}(\widehat{\mathbf{x}})$), and let us call the solution $\widehat{\mathbf{x}}$ *regular* if $\mathsf{B}_{*\mathscr{F}}$ is a full row rank matrix, and *range regular* if $\mathrm{Im}\mathsf{B} = \mathrm{Im}\mathsf{B}_{*\mathscr{F}}$. It is easy to check that the regular or range regular solution of (9.1) satisfies the Abadie constraint qualification.

Theorem 9.2 *Let $\{\mathbf{x}^k\}$, $\{\boldsymbol{\lambda}^k\}$ be generated by Algorithm 9.1 for the solution of (9.1) with $\eta > 0$, $\beta > 1$, $M > 0$, $\rho > 0$, and $\boldsymbol{\lambda}^0 \in \mathbb{R}^m$. Then the following statements hold.*
(i) The sequence $\{\mathbf{x}^k\}$ converges to the solution $\widehat{\mathbf{x}}$ of (9.1).
(ii) If the solution $\widehat{\mathbf{x}}$ of (9.1) is regular, then $\{\boldsymbol{\lambda}^k\}$ converges to a uniquely determined vector $\widehat{\boldsymbol{\lambda}}_E$ of Lagrange multipliers for the equality constraints of (9.1).
(iii) If the solution $\widehat{\mathbf{x}}$ of (9.1) is range regular, then $\{\boldsymbol{\lambda}^k\}$ converges to

$$\overline{\boldsymbol{\lambda}} = \boldsymbol{\lambda}_{LS} + (\mathsf{I} - \mathsf{P})\boldsymbol{\lambda}^0,$$

where P *is the orthogonal projector onto* $\operatorname{Im}\mathsf{B} = \operatorname{Im}\mathsf{B}_{*\mathscr{F}}$ *and* $\boldsymbol{\lambda}_{\mathrm{LS}}$ *is the least square Lagrange multiplier for the equality constraints of the solution of (9.1).*

Proof (i) Since the iterates $\mathbf{x}^k$ are bounded due to Lemma 9.1, it follows that there is a cluster point $\overline{\mathbf{x}}$ of $\{\mathbf{x}^k\}$ and $\mathscr{K} \subseteq \mathbb{N}$ such that

$$\lim_{k\to\infty} \{\mathbf{x}^k\}_{k\in\mathscr{K}} = \overline{\mathbf{x}}.$$

Moreover, since $\mathbf{x}^k \in \Omega_S$ and by (9.26)

$$\lim_{k\to\infty} \|\mathsf{B}\mathbf{x}^k\| = 0,$$

it follows that $\mathsf{B}\overline{\mathbf{x}} = \mathbf{o}$, $\overline{\mathbf{x}} \in \Omega_{SE}$, and $f(\overline{\mathbf{x}}) \geq f(\widehat{\mathbf{x}})$.

To show that $\overline{\mathbf{x}}$ solves (9.1), let us denote

$$\mathbf{d} = \widehat{\mathbf{x}} - \overline{\mathbf{x}}, \quad \alpha_k = \max\{\alpha \in [0,1] : \mathbf{x}^k + \alpha\mathbf{d} \in \Omega_S\}, \quad \mathbf{d}^k = \alpha_k\mathbf{d},$$

so that

$$\mathsf{B}\mathbf{d} = \mathsf{B}\mathbf{d}^k = \mathbf{o}, \quad \lim_{k\to\infty} \mathbf{d}^k = \mathbf{d}, \quad \lim_{k\to\infty} \mathbf{x}^k + \mathbf{d}^k = \widehat{\mathbf{x}},$$

and

$$L(\mathbf{x}^k + \mathbf{d}, \boldsymbol{\lambda}^k, \rho) - L(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho) = f(\mathbf{x}^k + \mathbf{d}) - f(\mathbf{x}^k).$$

Moreover, if we denote

$$\hat{\mathbf{x}}^k = \arg\min_{\mathbf{x}\in\Omega_S} L(\mathbf{x}, \boldsymbol{\lambda}^k, \rho),$$

then by Lemma 7.2 and the definition of $\mathbf{x}^k$ in Step 1 of SMALSE-M

$$0 \leq L(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho) - L(\hat{\mathbf{x}}^k, \boldsymbol{\lambda}^k, \rho) \leq \frac{1}{2\lambda_{\min}} \|\mathbf{g}^P(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho)\|^2 \leq \frac{M_0^2}{2\lambda_{\min}} \|\mathsf{B}\mathbf{x}^k\|^2.$$

Using the above relations, we get

$$\begin{aligned} 0 &\leq L(\mathbf{x}^k + \mathbf{d}^k, \boldsymbol{\lambda}^k, \rho) - L(\hat{\mathbf{x}}^k, \boldsymbol{\lambda}^k, \rho) \\ &= L(\mathbf{x}^k + \mathbf{d}^k, \boldsymbol{\lambda}^k, \rho) - L(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho) + L(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho) - L(\hat{\mathbf{x}}^k, \boldsymbol{\lambda}^k, \rho) \\ &\leq f(\mathbf{x}^k + \mathbf{d}^k) - f(\mathbf{x}^k) + \frac{M_0^2}{2\lambda_{\min}} \|\mathsf{B}\mathbf{x}^k\|^2. \end{aligned}$$

The continuity of f implies

$$0 \leq \lim_{k\to\infty} \left(f(\mathbf{x}^k + \mathbf{d}^k) - f(\mathbf{x}^k) + \frac{M_0^2}{2\lambda_{\min}} \|\mathsf{B}\mathbf{x}^k\|^2 \right) = f(\widehat{\mathbf{x}}) - f(\overline{\mathbf{x}}).$$

It follows that $\overline{\mathbf{x}}$ solves (9.1). The solution $\widehat{\mathbf{x}}$ of (9.1) being unique, it follows that $\mathbf{x}^k$ converges to $\overline{\mathbf{x}} = \widehat{\mathbf{x}}$.
(ii) Let us denote $\mathscr{F} = \widetilde{\mathscr{F}}(\widehat{\mathbf{x}})$ and $\mathsf{H} = \mathsf{A} + \rho \mathsf{B}^T \mathsf{B}$, so by the assumptions there is a unique Lagrange multiplier $\widehat{\boldsymbol{\lambda}}$ such that

$$[\mathsf{H}\widehat{\mathbf{x}} - \mathbf{b} + \mathsf{B}^T\widehat{\boldsymbol{\lambda}}]_{\mathscr{F}} = \mathbf{o}. \tag{9.30}$$

Since we have just proved that $\{\mathbf{x}^k\}$ converges to $\widehat{\mathbf{x}}$, there is k_1 such that $\mathscr{F} \subseteq \mathscr{F}\{\mathbf{x}^k\}$ for $k \geq k_1$ and

$$\mathbf{g}_{\mathscr{F}}(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho) = \mathsf{H}_{\mathscr{F}*}\mathbf{x}^k - \mathbf{b}_{\mathscr{F}} + \mathsf{B}_{*\mathscr{F}}^T\boldsymbol{\lambda}^k$$

converges to zero. It follows that the sequence

$$\mathsf{B}_{*\mathscr{F}}^T\boldsymbol{\lambda}^k = \mathbf{b}_{\mathscr{F}} - \mathsf{H}_{\mathscr{F}*}\mathbf{x}^k + \mathbf{g}_{\mathscr{F}}(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho)$$

is bounded. Since the largest nonzero singular value $\overline{\sigma}_{\min}$ of $\mathsf{B}_{*\mathscr{F}}$ satisfies

$$\overline{\sigma}_{\min}\|\boldsymbol{\lambda}^k\| \leq \|\mathsf{B}_{*\mathscr{F}}^T\boldsymbol{\lambda}^k\|,$$

it follows that there is a cluster point $\overline{\boldsymbol{\lambda}}$ of $\boldsymbol{\lambda}^k$. Moreover, $\overline{\boldsymbol{\lambda}}$ satisfies

$$[\mathsf{H}\widehat{\mathbf{x}} - \mathbf{b} + \mathsf{B}^T\overline{\boldsymbol{\lambda}}]_{\mathscr{F}} = \mathbf{o}.$$

After comparing the last relation with (9.30) and using the assumptions, we conclude that $\overline{\boldsymbol{\lambda}} = \widehat{\boldsymbol{\lambda}}$ and $\boldsymbol{\lambda}^k$ converges to $\widehat{\boldsymbol{\lambda}}$.
(iii) Let us assume that the solution $\widehat{\mathbf{x}}$ of (9.1) is only range regular, let $\overline{\boldsymbol{\lambda}}$ denote any vector of Lagrange multipliers for (9.1), and let $\mathsf{Q} = \mathsf{I} - \mathsf{P}$ denote the orthogonal projector onto $\operatorname{Ker}\mathsf{B}^T = \operatorname{Ker}\mathsf{B}_{*\mathscr{F}}^T$. Using $\mathsf{P} + \mathsf{Q} = \mathsf{I}$, $\mathsf{B}^T\mathsf{Q} = \mathsf{O}$, and (2.34), we get

$$\begin{aligned}\|\mathsf{B}_{*\mathscr{F}}^T(\boldsymbol{\lambda}^k - \overline{\boldsymbol{\lambda}})\| &= \|\mathsf{B}_{*\mathscr{F}}^T(\mathsf{P} + \mathsf{Q})(\boldsymbol{\lambda}^k - \overline{\boldsymbol{\lambda}})\| = \|\mathsf{B}_{*\mathscr{F}}^T(\mathsf{P}\boldsymbol{\lambda}^k - \mathsf{P}\overline{\boldsymbol{\lambda}})\| \\ &\geq \overline{\sigma}_{\min}\|\mathsf{P}\boldsymbol{\lambda}^k - \mathsf{P}\overline{\boldsymbol{\lambda}}\|.\end{aligned}$$

Using the arguments from the proof of (ii), we get that the left hand side of the above relation converges to zero, so $\mathsf{P}\boldsymbol{\lambda}^k$ converges to $\mathsf{P}\overline{\boldsymbol{\lambda}}$. Since

$$\boldsymbol{\lambda}^k = \boldsymbol{\lambda}^0 + \rho\mathsf{B}\mathbf{x}^0 + \cdots + \rho_k\mathsf{B}\mathbf{x}^k$$

with $\mathsf{B}\mathbf{x}^k \in \operatorname{Im}\mathsf{B}$, we get

$$\boldsymbol{\lambda}^k = (\mathsf{P} + \mathsf{Q})\boldsymbol{\lambda}^k = \mathsf{Q}\boldsymbol{\lambda}^0 + \mathsf{P}\boldsymbol{\lambda}^k.$$

Observing that $\overline{\boldsymbol{\lambda}} = \boldsymbol{\lambda}_{\mathrm{LS}} + \mathsf{Q}\boldsymbol{\lambda}^0$ is a Lagrange multiplier for (9.1) and $\mathsf{P}\overline{\boldsymbol{\lambda}} = \boldsymbol{\lambda}_{\mathrm{LS}}$, we get

$$\|\boldsymbol{\lambda}^k - \overline{\boldsymbol{\lambda}}\| = \|\mathsf{Q}\boldsymbol{\lambda}^0 + \mathsf{P}\boldsymbol{\lambda}^k - \left(\boldsymbol{\lambda}_{\mathrm{LS}} + \mathsf{Q}\boldsymbol{\lambda}^0\right)\| = \|\mathsf{P}\boldsymbol{\lambda}^k - \mathsf{P}\overline{\boldsymbol{\lambda}}\|.$$

Since the right-hand side converges to zero, we conclude that $\boldsymbol{\lambda}^k$ converges to $\overline{\boldsymbol{\lambda}}$, which completes the proof of (iii). □

9.8 Optimality of the Outer Loop

Theorem 9.1 suggests that it is possible to give an independent of B upper bound on the number of outer iterations of Algorithm 9.1 (SMALSE-M) that are necessary to achieve a prescribed feasibility error for a class of problems like (9.1). To present explicitly this new feature of SMALSE-M, at least as compared to the related algorithms [7], let $\mathscr{T}$ denote any set of indices and let for any $t \in \mathscr{T}$ be defined the problem

$$\text{minimize} \;\; f_t(\mathbf{x}) \;\; \text{s.t.} \;\; \mathbf{x} \in \Omega_{SE}^t, \quad f_t(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \mathsf{A}_t \mathbf{x} - \mathbf{b}_t^T \mathbf{x}, \tag{9.31}$$

where

$$\Omega_{SE}^t = \{\mathbf{x} \in \mathbb{R}^{n_t} : \mathsf{B}_t \mathbf{x} = \mathbf{o} \;\text{ and }\; \mathbf{x} \in \Omega_S^t\}, \quad \Omega_S^t = \{h_i^t(\mathbf{x}_i) \le 0, \; i \in \mathscr{I}^t\},$$

h_i^t defines the bound, spherical, or elliptic constraints, $\mathsf{A}_t \in \mathbb{R}^{n_t \times n_t}$ denotes an SPD matrix, $\mathsf{B}_t \in \mathbb{R}^{m_t \times n_t}$, and $\mathbf{b}_t \in \mathbb{R}^{n_t}$. Our optimality result reads as follows.

Theorem 9.3 *Let $\{\mathbf{x}_t^k\}$, $\{\boldsymbol{\lambda}_t^k\}$, and $\{M_{t,k}\}$ be generated by Algorithm 9.1 for (9.31) with*

$$0 < \eta_t \le \|\mathbf{b}_t\|, \;\; 0 < \beta < 1, \;\; M_{t,0} = M_0 > 0, \;\; \rho > 0, \;\; \text{and} \;\; \boldsymbol{\lambda}_t^0 = \mathbf{o}.$$

Let $\mathbf{o} \in \Omega_{SE}^t$ and let there be an $a_{\min} > 0$ such that the least eigenvalue $\lambda_{\min}(\mathsf{A}_t)$ of the Hessian A_t of the quadratic function f_t satisfies

$$\lambda_{\min}(\mathsf{A}_t) \ge a_{\min}, \quad t \in \mathscr{T}.$$

Then for each $\varepsilon > 0$ there are indices k_t, $t \in \mathscr{T}$, such that

$$k_t \le a/\varepsilon^2 + 1 \tag{9.32}$$

and $\mathbf{x}_t^{k_t}$ is an approximate solution of (9.31) satisfying

$$\|\mathsf{B}_t \mathbf{x}_t^{k_t}\| \le \varepsilon \|\mathbf{b}_t\|. \tag{9.33}$$

Proof First notice that for any index j

$$\frac{\rho j}{2} \min\{\|\mathsf{B}_t \mathbf{x}_t^i\|^2 : \; i = 1, \ldots, j\} \le \sum_{i=1}^{j} \frac{\rho}{2} \|\mathsf{B}_t \mathbf{x}_t^i\|^2 \le \sum_{i=1}^{\infty} \frac{\rho}{2} \|\mathsf{B}_t \mathbf{x}_t^i\|^2. \tag{9.34}$$

Denoting by $L_t(\mathbf{x}, \boldsymbol{\lambda}, \rho)$ the augmented Lagrangian for problem (9.31), we get for any $\mathbf{x} \in \mathbb{R}^{n_t}$ and $\rho \geq 0$

$$L_t(\mathbf{x}, \mathbf{o}, \rho) = \frac{1}{2}\mathbf{x}^T(\mathsf{A}_t + \rho \mathsf{B}_t^T \mathsf{B}_t)\mathbf{x} - \mathbf{b}_t^T \mathbf{x} \geq \frac{1}{2} a_{\min} \|\mathbf{x}\|^2 - \|\mathbf{b}_t\| \|\mathbf{x}\| \geq -\frac{\|\mathbf{b}_t\|^2}{2a_{\min}}.$$

If we substitute this inequality and $\mathbf{z}_0 = \mathbf{o}$ into (9.25) and use the assumption $\|\mathbf{b}_t\| \geq \eta_t$, we get

$$\sum_{i=1}^{\infty} \frac{\rho}{2} \|\mathsf{B}_t \mathbf{x}_t^i\|^2 \leq \frac{\|\mathbf{b}_t\|^2}{2a_{\min}} + (1+p)\frac{\eta^2}{2a_{\min}} \leq \frac{(2+p)\|\mathbf{b}_t\|^2}{2a_{\min}}, \tag{9.35}$$

where $p \geq 0$ denotes the smallest integer such that $\rho \geq \beta^{2p} M_0^2 / a_{\min}$. Using (9.34) and (9.35), we get

$$\frac{\rho j}{2} \min\{\|\mathsf{B}_t \mathbf{x}_t^i\|^2 : i = 1, \ldots, j\} \leq \frac{(2+p)}{2a_{\min}\varepsilon^2} \varepsilon^2 \|\mathbf{b}_t\|^2.$$

Let us now denote

$$a = (2+p)/(a_{\min}\rho)$$

and take for j the least integer which satisfies $a/j \leq \varepsilon^2$, so that

$$a/\varepsilon^2 \leq j \leq a/\varepsilon^2 + 1. \tag{9.36}$$

Denoting for any $t \in \mathscr{T}$

$$k_t = \arg\min\{\|\mathsf{B}_t \mathbf{x}_t^i\| : i = 1, \ldots, j\},$$

we can use (9.36) with simple manipulations to obtain

$$\|\mathsf{B}_t \mathbf{x}_t^{k_t}\|^2 = \min\{\|\mathsf{B}_t \mathbf{x}_t^i\|^2 : i = 1, \ldots, j\} \leq \frac{a}{j\varepsilon^2} \varepsilon^2 \|\mathbf{b}_t\|^2 \leq \varepsilon^2 \|\mathbf{b}_t\|^2. \qquad \square$$

9.9 Optimality of the Inner Loop

We need the following simple lemma to prove the optimality of the inner loop (Step 1) implemented by the MPGP algorithm.

Lemma 9.4 *Let* $\{\mathbf{x}^k\}$ *and* $\{\boldsymbol{\lambda}^k\}$ *be generated by Algorithm 9.1 for the solution of (9.1) with* $\eta > 0$, $0 < \beta < 1$, $M > 0$, $\rho > 0$, *and* $\boldsymbol{\lambda}^0 \in \mathbb{R}^m$. *Let* $0 < a_{\min} \leq \lambda_{\min}(\mathsf{A})$, *where* $\lambda_{\min}(\mathsf{A})$ *denotes the least eigenvalue of* A.

Then for any $k \geq 0$

$$L(\mathbf{x}^k, \lambda^{k+1}, \rho) - L(\mathbf{x}^{k+1}, \boldsymbol{\lambda}^{k+1}, \rho) \leq \frac{\eta^2}{2a_{\min}}. \tag{9.37}$$

Proof Let us denote

$$\hat{\mathbf{x}}^{k+1} = \arg \min_{\mathbf{x} \in \Omega_S} L(\mathbf{x}, \boldsymbol{\lambda}^{k+1}, \rho).$$

Using Lemma 7.2 and the definition of Step 1 of SMALSE-M, we get

$$\begin{aligned} L(\mathbf{x}^k, \boldsymbol{\lambda}^{k+1}, \rho) - L(\mathbf{x}^{k+1}, \boldsymbol{\lambda}^{k+1}, \rho) &\leq L(\mathbf{x}^k, \boldsymbol{\lambda}^{k+1}, \rho) - L(\hat{\mathbf{x}}^{k+1}, \boldsymbol{\lambda}^{k+1}, \rho) \\ &\leq \frac{1}{2a_{\min}} \|\mathbf{g}^P\|^2 \leq \frac{\eta^2}{2a_{\min}}. \end{aligned}$$

□

Now we are ready to prove the main result of this chapter, the optimality of Algorithm 9.1 (SMALSE-M) in terms of matrix–vector multiplications, provided Step 1 is implemented by Algorithm 7.2 (MPGP) or any other, possibly more specialized algorithm for the solution of separable constraints with R-linear rate of convergence of the norm of the projected gradient, such as MPRGP described in Chap. 8.

Theorem 9.4 *Let*

$$0 < a_{\min} < a_{\max}, \quad 0 < c_{\max}, \quad \textit{and} \quad \varepsilon > 0$$

be given constants and let the class of problems (9.31) satisfy

$$a_{\min} \leq \lambda_{\min}(\mathsf{A}_t) \leq \lambda_{\max}(\mathsf{A}_t) \leq a_{\max} \ \textit{and} \ \|\mathsf{B}_t\| \leq c_{\max}. \tag{9.38}$$

Let $\{\mathbf{x}_t^k\}, \{\boldsymbol{\lambda}_t^k\}$, *and* $\{M_{t,k}\}$ *be generated by Algorithm 9.1 (SMALSE-M) for (9.31) with*

$$\|\mathbf{b}_t\| \geq \eta_t > 0, \quad 0 < \beta < 1, \quad M_{t,0} = M_0 > 0, \quad \rho > 0, \quad \textit{and} \quad \boldsymbol{\lambda}_t^0 = \mathbf{o}.$$

Let Step 1 of Algorithm 9.1 be implemented by Algorithm 7.2 (MPGP) with the parameters $\Gamma > 0$ *and* $\alpha \in (0, 2(a_{\max} + \rho c_{\max}^2)^{-1}]$ *to generate the iterates* $\mathbf{x}_t^{k,0}, \mathbf{x}_t^{k,1}, \ldots, \mathbf{x}_t^{k,l} = \mathbf{x}_t^k$ *for the solution of (9.31) starting from* $\mathbf{x}_t^{k,0} = \mathbf{x}_t^{k-1}$ *with* $\mathbf{x}_t^{-1} = \mathbf{o}$, *where* $l = l_{t,k}$ *is the first index satisfying*

$$\|g^P(\mathbf{x}_t^{k,l}, \boldsymbol{\lambda}_t^k, \rho)\| \leq M_{t,\ell} \|\mathsf{B}_t \mathbf{x}_t^{k,\ell}\|. \tag{9.39}$$

Then Algorithm 9.1 generates an approximate solution $\mathbf{x}_t^{k_t}$ *of any problem (9.31) which satisfies*

$$\|\mathbf{g}^P(\mathbf{x}_t^{k_t}, \boldsymbol{\lambda}_t^{k_t}, \rho)\| \leq \varepsilon \|\mathbf{b}_t\| \ \textit{and} \ \|\mathsf{B}_t \mathbf{x}_t^{k_t}\| \leq \varepsilon \|\mathbf{b}_t\| \tag{9.40}$$

at $O(1)$ matrix–vector multiplications by the Hessian of the augmented Lagrangian for (9.31).

Proof Let $t \in \mathscr{T}$ be fixed and let us denote by $L_t(\mathbf{x}, \boldsymbol{\lambda}, \rho)$ the augmented Lagrangian for problem (9.31). Then by (9.37) and the assumption $\eta_t \le \|\mathbf{b}_t\|$

$$L_t(\mathbf{x}_t^{k-1}, \boldsymbol{\lambda}_t^k, \rho) - L_t(\mathbf{x}_t^k, \boldsymbol{\lambda}_t^k, \rho) \le \frac{\eta_t^2}{2a_{\min}} \le \frac{\|\mathbf{b}_t\|^2}{2a_{\min}}.$$

Since the minimizer $\overline{\mathbf{x}}_t^k$ of $L_t(\mathbf{x}, \boldsymbol{\lambda}_t^k, \rho)$ subject to $\mathbf{x} \in \Omega_S^t$ satisfies (9.8) and is a possible choice for $\mathbf{x}_t^k$, it follows that

$$L_t(\mathbf{x}_t^{k-1}, \boldsymbol{\lambda}_t^k, \rho) - L_t(\overline{\mathbf{x}}_t^k, \boldsymbol{\lambda}_t^k, \rho) \le \frac{\|\mathbf{b}_t\|^2}{2a_{\min}}. \tag{9.41}$$

Using Theorem 7.3, we get that Algorithm 7.2 (MPGP) used to implement Step 1 of Algorithm 9.1 (SMALSE-M) starting from $\mathbf{x}_t^{k,0} = \mathbf{x}_t^{k-1}$ generates $\mathbf{x}_t^{k,l}$ satisfying

$$\|g_t^P(\mathbf{x}_t^{k,l}, \boldsymbol{\lambda}_t^k, \rho)\|^2 \le a_1 \eta_\Gamma^l \big(L_t(\mathbf{x}_t^{k-1}, \boldsymbol{\lambda}_t^k, \rho) - L_t(\overline{\mathbf{x}}_t^k, \boldsymbol{\lambda}_t^k, \rho)\big) \le a_1 \frac{\|\mathbf{b}_t\|^2}{2a_{\min}} \eta_\Gamma^l,$$

where a_1 and $\eta = \eta(\delta, \alpha) < 1$ are the constants specified for the class of problems (9.31) in Theorem 7.3. It simply follows by the inner stop rule (9.39) that the number l of the inner iterations in Step 1 is uniformly bounded by an index $l_{\max}$ which satisfies

$$a_1 \frac{\|\mathbf{b}_t\|^2}{2a_{\min}} \eta_\Gamma^{l_{\max}} \le M_{t,\ell_{\max}}^2 \varepsilon^2 \|\mathbf{b}_t\|^2.$$

Thus we have proved that Step 1 implemented by MPGP is completed in a uniformly bounded number of iterations of MPGP. Since each iteration of MPRGP requires at most two matrix–vector multiplications, it follows that Step 1 can be carried out for any instance of the class of problems (9.1) in a uniformly bounded number of matrix–vector multiplications.

To finish the proof, it is enough to combine this result with Theorem 9.3, in particular to carry out the outer iterations until

$$M_{t,k_t} \|\mathsf{B}\mathbf{x}^{k_t}\| \le \varepsilon \|\mathbf{b}\| \quad \text{and} \quad \|\mathsf{B}\mathbf{x}^{k_t}\| \le \varepsilon \|\mathbf{b}\|.$$

□

9.10 SMALBE for Bound and Equality Constrained QP Problems

We shall now consider a special case of problem (9.1), the minimization of a strictly convex quadratic function on a feasible set defined by the *bound and linear equality constraints*

$$\min_{\mathbf{x}\in\Omega_{BE}} f(\mathbf{x}), \quad \Omega_{BE} = \{\mathbf{x}\in\mathbb{R}^n : \mathsf{B}\mathbf{x} = \mathbf{o} \quad \text{and} \quad \mathbf{x} \geq \ell\}, \tag{9.42}$$

where $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathsf{A}\mathbf{x} - \mathbf{x}^T\mathbf{b}$, $\mathbf{b}\in\mathbb{R}^n$, A is an $n\times n$ SPD matrix, and $\mathsf{B}\in\mathbb{R}^{m\times n}$. We consider the same assumptions as above, in particular, $\Omega_{BE}\neq\emptyset$, $\mathsf{B}\neq\mathsf{O}$, and $\operatorname{Ker}\mathsf{B}\neq\{\mathbf{o}\}$. We admit dependent rows of B and $\ell_i = -\infty$.

The complete algorithm that we call SMALBE-M (Semi-Monotonic Augmented Lagrangians for Bound and Equality constraints) reads as follows.

Algorithm 9.2 Semi-monotonic augmented Lagrangians for bound and equality constrained QP problems (SMALBE-M).

Given an SPD matrix $\mathsf{A}\in\mathbb{R}^{n\times n}$, $\mathsf{B}\in\mathbb{R}^{m\times n}$, *n-vectors* $\mathbf{b}$, ℓ.
Step 0. {*Initialization.*}
Choose $\eta > 0$, $0 < \beta < 1$, $M_{-1} = 0$, $M_0 > 0$, $\rho > 0$, $\boldsymbol{\lambda}^0\in\mathbb{R}^m$
for $k = 0, 1, 2, \ldots$
Step 1. {*Inner iteration with adaptive precision control.*}
Find $\mathbf{x}^k \geq \ell$ *such that*

$$\|\mathbf{g}^P(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho)\| \leq \min\{M_k\|\mathsf{B}\mathbf{x}^k\|, \eta\} \tag{9.43}$$

Step 2. {*Updating the Lagrange multipliers.*}

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \rho\mathsf{B}\mathbf{x}^k \tag{9.44}$$

Step 3. {*Update M provided the increase of the Lagrangian is not sufficient.*}
if $M_k = M_{k-1}$ *and*

$$L(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho) < L(\mathbf{x}^{k-1}, \boldsymbol{\lambda}^{k-1}, \rho) + \frac{\rho}{2}\|\mathsf{B}\mathbf{x}^k\|^2 \tag{9.45}$$

$M_{k+1} = \beta M_k$
else
$M_{k+1} = M_k$
end else if
end for

In Step 1 we can use any algorithm for minimizing strictly convex quadratic functions subject to bound constraints as long as it guarantees the convergence of projected gradient to zero, such as the MPRGP algorithm of Chap. 8.

9.11 R-Linear Convergence of SMALBE-M

Our main optimality result, Theorem 9.4, guarantees that the number of iterations $k_t(\varepsilon)$ that are necessary to get an approximate solution of any problem from the class of problems (9.31) to the prescribed relative precision ε is uniformly bounded by

$$k(\varepsilon) = \max\{k_t(\varepsilon) : t \in \mathcal{T}\}.$$

Notice that $k(\varepsilon)$ does not depend on the matrices B_t which appear in the description of the feasible sets $\Omega_t \ldots$ a feature which is not obvious from the standard analysis of the Uzawa type methods even for linear problems. However, Theorem 9.3 yields only

$$k(\varepsilon) \lesssim \varepsilon^{-2},$$

i.e., we proved that there is $C > 0$ independent of B_t such that for any $\varepsilon > 0$

$$k(\varepsilon) \leq C\varepsilon^{-2},$$

which is very pessimistic and has never been observed in practice.

Here we report a stronger result concerning the convergence of a particular class of problem (9.42) defined by the bound and equality constraints, namely that SMALBE-M enjoys the R-linear convergence of feasibility errors in a later stage of computations, when the indices of the free and strongly active variables of the solution are identified, and show that there are $k_t(\varepsilon)$ and $\overline{k}_t$ such that for sufficiently small ε

$$k_t(\varepsilon) - \overline{k}_t \lesssim |\log(\varepsilon)|.$$

Notice that the iterates can remain nonlinear in any stage of the solution procedure, so the convergence analysis cannot be reduced to that for the equality constraints. The result does not assume independent equality constraints and remains valid even when there are some zero multipliers for active bound constraints.

We shall start with the following lemma which shows that the binding set $\widehat{\mathcal{B}} = \mathcal{B}(\widehat{\mathbf{x}})$ and the free set $\widehat{\mathcal{F}} = \mathcal{F}(\widehat{\mathbf{x}})$ are identified in a finite number of steps.

Lemma 9.5 *Let the matrices* A, B *and the vectors* $\mathbf{b}$, ℓ *be those from the definition of problem (9.42). Let* $\mathbf{x}^k$ *and* $\boldsymbol{\lambda}^k$ *be generated by the SMALBE-M algorithm for the solution of (9.42) with the regularization parameter* $\rho > 0$. *Then there is* k_1 *such that for* $k \geq k_1$

$$\widehat{\mathcal{F}} = \mathcal{F}(\widehat{\mathbf{x}}) \subseteq \mathcal{F}(x^k) \quad \text{and} \quad \widehat{\mathcal{B}} = \mathcal{B}(\widehat{\mathbf{x}}) \subseteq \mathcal{B}(\mathbf{x}^k). \tag{9.46}$$

Proof First observe that $\mathbf{g}^P(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho)$ converges to zero by Theorem 9.3. If $\widehat{\mathcal{B}} = \emptyset$ or $\widehat{\mathcal{F}} = \emptyset$, then the statements of our lemma concerning these sets are valid trivially. Hence we assume that they are both nonempty and denote

$$\delta_{\widehat{\mathcal{F}}} = \min\{\widehat{x}_i - \ell_i : i \in \widehat{\mathcal{F}}\}, \quad \delta_{\widehat{\mathcal{B}}} = \min\{\widehat{g}_i : i \in \widehat{\mathcal{B}}\}.$$

Since $\mathbf{x}^k$ converges to $\widehat{\mathbf{x}}$, there is k' such that for $k \geq k'$ and $i \in \widehat{\mathscr{F}}$

$$\ell_i < \widehat{x}_i - \frac{1}{2}\delta_{\widehat{\mathscr{F}}} \leq x_i^k,$$

i.e., $i \in \mathscr{F}(\mathbf{x}^k)$.

To prove the second relation of (9.46), observe that $\mathbf{g}$ is a continuous function, so $\mathbf{g}^k = \mathbf{g}(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho)$ converges to $\widehat{\mathbf{g}} = \mathbf{g}(\widehat{\mathbf{x}}, \overline{\boldsymbol{\lambda}}, \rho)$. It follows that if $i \in \widehat{\mathscr{B}}$, then there is k'' such that for $k \geq k''$

$$g_i(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho) = g_i^k \geq \frac{1}{2}\delta_{\widehat{\mathscr{K}}}.$$

Moreover, since $\mathbf{g}^P(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho)$ converges to zero by the assumption, it follows that there is $k''' \geq k''$ such that for $k \geq k'''$ holds $x_i^k = \ell_i$, i.e., $i \in \mathscr{B}(x^k)$. We have thus proved the second relation of (9.46) with $k_1 = \max\{k', k'''\}$. □

Now we are ready to formulate the results showing that the convergence is R-linear after the free and strong active constraints of the solution are identified.

Theorem 9.5 *Let $\{\mathbf{x}^k\}$, $\{\boldsymbol{\lambda}^k\}$, and $\{M_k\}$ be generated by SMALBE-M for the solution to (9.42) with $\eta > 0$, $0 < \beta < 1$, $M_0 > 0$, $\rho > 0$, and $\boldsymbol{\lambda}^0 \in \mathbb{R}^m$. Let the solution $\widehat{\mathbf{x}}$ of (9.42) be range regular and let k_0 and k_1 be those of Theorem 9.1 and Lemma 9.5, so that for $\overline{k} = \max\{k_0, k_1\}$*

$$M_{\overline{k}} = M_{\overline{k}+1} = M_{\overline{k}+2} = \cdots \quad \text{and} \quad \mathscr{F}(\widehat{\mathbf{x}}) = \mathscr{F}(\mathbf{x}^{\overline{k}}) = \mathscr{F}(\mathbf{x}^{\overline{k}+1}) = \mathscr{F}(\mathbf{x}^{\overline{k}+2}) = \cdots.$$

Let $\widehat{\sigma}_{\min}$ denote the smallest nonzero singular value of $\mathsf{B}_{*\widehat{\mathscr{F}}}$, let $\mathsf{H} = \mathsf{A} + \rho\mathsf{B}^T\mathsf{B}$, and denote

$$C_1 = M_{k_0}\frac{\kappa(\mathsf{H})+1}{\lambda_{\min}(\mathsf{H})} + \frac{\kappa(A^\rho)}{\widehat{\sigma}_{\min}}, \quad C_2 = \frac{M_{k_0}\kappa(\mathsf{A}^\rho) + \widehat{\sigma}_{\min}^{-1}\|\mathsf{H}\|}{\widehat{\sigma}_{\min}} \quad C = \frac{2C_2}{\rho}.$$

Then the following relations hold:
(i) For any $k \geq \overline{k}$

$$\|\mathsf{B}\mathbf{x}^k\|^2 \leq (C+1)\left(\frac{C}{C+1}\right)^{k-\overline{k}}\|\mathsf{B}\mathbf{x}^{\overline{k}}\|^2. \tag{9.47}$$

(ii) For any $k \geq \overline{k}$

$$\|\mathbf{x}^k - \widehat{\mathbf{x}}\|^2 \leq C_1(C+1)\left(\frac{C}{C+1}\right)^{k-\overline{k}}\|\mathsf{B}\mathbf{x}^{\overline{k}}\|^2. \tag{9.48}$$

(iii) If $\boldsymbol{\lambda}^0 \in \mathrm{Im}\mathsf{B}$*, then for any* $k \geq \overline{k}$

$$\|\boldsymbol{\lambda}^k - \boldsymbol{\lambda}_{LS}\|^2 \leq C_2(C+1)\left(\frac{C}{C+1}\right)^{k-\overline{k}} \|\mathsf{B}\mathbf{x}^{\overline{k}}\|^2. \tag{9.49}$$

Proof See [10, Theorem 4.4].

9.12 SMALSE-Mw

The algorithm SMALSE-Mw that we develop here is a modification of SMALSE-M which can cope with a high sensitivity of the projected gradient when the curvature of the boundary of a feasible set is strong, as typically happens when the inequality constraints are elliptic as in the algorithms for contact problems with orthotropic friction. The difficulties with the curvature are resolved by using the *reduced projected gradient* $\widetilde{\mathbf{g}}_\alpha^P$ (see Sect. 7.2). The SMALSE-Mw algorithm reads as follows.

Algorithm 9.3 Semimonotonic augmented Lagrangians for separable and equality constrained QCQP problems (SMALSE-Mw).

Given an SPD matrix $\mathsf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, $\mathsf{B} \in \mathbb{R}^{m\times n}$, *constraints* $\mathbf{h}$.
Step 0. {*Initialization.*}
Choose $\eta > 0$, $0 < \beta < 1$, $M_{-1} = 0$, $M_0 > 0$, $\rho > 0$, $\boldsymbol{\lambda}^0 \in \mathbb{R}^m$
for $k = 0, 1, 2, \ldots$
Step 1. {*Inner iteration with adaptive precision control.*}
Find $\mathbf{x}^k \in \Omega_S$ *such that*

$$\|\widetilde{\mathbf{g}}^P(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho)\| \leq \min\{M_k\|\mathsf{B}\mathbf{x}^k\|, \eta\} \tag{9.50}$$

Step 2. {*Updating the Lagrange multipliers.*}

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \rho\mathsf{B}\mathbf{x}^k \tag{9.51}$$

Step 3. {*Update M provided the increase of the Lagrangian is not sufficient.*}
if $M_k = M_{k-1}$ *and*

$$L(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho) < L(\mathbf{x}^{k-1}, \boldsymbol{\lambda}^{k-1}, \rho) + \frac{\rho}{2}\|\mathsf{B}\mathbf{x}^k\|^2 \tag{9.52}$$

$M_{k+1} = \beta M_k$
else
$M_{k+1} = M_k$
end else if
end for

In Step 1 we can use any algorithm for minimizing the strictly convex quadratic function subject to separable constraints as long as it guarantees the convergence of reduced projected gradients to zero. Since Theorem 7.1 guarantees that for any $\boldsymbol{\lambda}^k \in \mathbb{R}^m$ and $\rho \geq 0$, there is a constant $C > 0$ such that for any $\mathbf{x} \in \Omega_S$

$$\|\tilde{\mathbf{g}}_\alpha^P(\mathbf{x}, \boldsymbol{\lambda}^k, \rho)\| \leq \|\mathbf{g}^P(\mathbf{x}, \boldsymbol{\lambda}^k, \rho)\| \leq C\|\tilde{\mathbf{g}}_\alpha^P(\mathbf{x}, \boldsymbol{\lambda}^k, \rho)\|, \tag{9.53}$$

it follows that we can use the MPGP algorithm of Sect. 7.3.

The next lemma shows that Algorithm 9.3 is well defined, that is, any algorithm for the solution of auxiliary problems required in Step 1 that guarantees the convergence of the reduced projected gradient to zero generates either a feasible $\mathbf{x}^k$ which satisfies (9.50) in a finite number of steps or approximations which converge to the solution of (9.1).

Lemma 9.6 *Let $M > 0$, $\boldsymbol{\lambda} \in \mathbb{R}^m$, $\alpha \in (0, 2(\|\mathsf{A}\| + \rho\|\mathsf{B}\|^2)^{-1}\|)$, $\eta > 0$, and $\rho \geq 0$ be given. Let $\{\mathbf{y}^k\} \in \Omega_S$ denote any sequence such that*

$$\widehat{\mathbf{x}} = \lim_{k\to\infty} \mathbf{y}^k = \arg\min_{\mathbf{y}\in\Omega_S} L(\mathbf{y}, \boldsymbol{\lambda}, \rho)$$

and $\tilde{\mathbf{g}}_\alpha^P(\mathbf{y}^k, \boldsymbol{\lambda}, \rho)$ converges to the zero vector. Then $\{\mathbf{y}^k\}$ either converges to the unique solution $\widehat{\mathbf{x}}$ of problem (9.1), or there is an index k such that

$$\|\tilde{\mathbf{g}}_\alpha^P(\mathbf{y}^k, \boldsymbol{\lambda}, \rho)\| \leq \min\{M\|\mathsf{B}\mathbf{y}^k\|, \eta\}. \tag{9.54}$$

Proof If (9.54) does not hold for any k, then $\|\tilde{\mathbf{g}}_\alpha^P(\mathbf{y}^k, \boldsymbol{\lambda}, \rho)\| > M\|\mathsf{B}\mathbf{y}^k\|$ for any k. Since $\tilde{\mathbf{g}}^P(\mathbf{y}_\alpha^k, \boldsymbol{\lambda}, \rho)$ converges to the zero vector by the assumption, it follows that $\|\mathsf{B}\mathbf{y}^k\|$ converges to zero. Thus $\mathsf{B}\widehat{\mathbf{y}} = \mathbf{o}$ and using the assumptions and (9.5), we get

$$\tilde{\mathbf{g}}_\alpha^P(\widehat{\mathbf{y}}) = \mathbf{g}^P(\widehat{\mathbf{y}}, \boldsymbol{\lambda}, \rho) = \mathbf{o}.$$

It follows that $\widehat{\mathbf{y}}$ satisfies the KKT conditions (9.5) and $\widehat{\mathbf{y}} = \widehat{\mathbf{x}}$. □

Inequality (9.53) guarantees that $\mathbf{x}^k$ satisfies

$$\|\mathbf{g}^P(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho)\| \leq \min\{CM_k\|\mathsf{B}\mathbf{x}^k\|, \eta\}, \tag{9.55}$$

so $\mathbf{x}^k$ can be considered as an iterate of SMALSE-M for problem (9.1). This is sufficient to guarantee the convergence of SMALSE-Mw. However, C in (9.55) depends on $\boldsymbol{\lambda}^k$, so the analysis used above is not sufficient to obtain optimality results for SMALSE-Mw. In spite of this, it turns out that SMALSE-Mw is an effective algorithms for the solution of problems with linear equality and separable inequality constrains with strong curvature, such as those arising in the solution of contact problems with orthotropic friction. Since $\boldsymbol{\lambda}^k$ changes slowly, it is not surprising that we observe in our experiments a kind of optimal performance.

9.13 Solution of More General Problems

If A is positive definite only on the kernel of B, then we can use a suitable penalization to reduce such problem to the strictly convex one. Using Lemma [11, Lemma 1.3], it is easy to see that there is $\overline{\rho} > 0$ such that $\mathsf{A} + \overline{\rho}\mathsf{B}^T\mathsf{B}$ is positive definite, so that we can apply our SMALBE-M algorithm to the equivalent penalized problem

$$\min_{\mathbf{x}\in\Omega_{BE}} f_{\overline{\rho}}(\mathbf{x}), \tag{9.56}$$

where

$$f_{\overline{\rho}}(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T(\mathsf{A} + \overline{\rho}\mathsf{B}^T\mathsf{B})\mathbf{x} - \mathbf{b}^T\mathbf{x}.$$

If A is an SPS matrix which is positive definite on the kernel of B, which is typical for the dual formulation of the problems arising from the discretization of contact problems, then we can use any $\overline{\rho} > 0$, typically $\overline{\rho} = \|\mathsf{A}\|$.

9.14 Implementation

Let us give here a few hints that can be helpful for an effective implementation of SMALSE-M (SMALBE-M) and SMALSE-Mw with the inner loop implemented by MPGP (MPRGP).

Before applying the algorithms presented to the problems with a well-conditioned Hessian A, we strongly recommend to rescale the equality constraints so that $\|\mathsf{A}\| \approx \|\mathsf{B}\|$. Taking into account the estimate of the rate of convergence in Theorem 9.5, it is also useful to orthonormalize or at least normalize the constraints.

A stopping criterion should be added not only after Step 1 but also into the procedure which generates $\mathbf{x}^k$ in Step 1. In our experiments, we use

$$\|\nabla L(\mathbf{x}^k, \boldsymbol{\lambda}^k, \rho)\| \le \varepsilon_g\|\mathbf{b}\| \ \text{ and } \ \|\mathsf{B}\mathbf{x}^k - \mathbf{c}\| \le \varepsilon_f\|\mathbf{b}\|. \tag{9.57}$$

The relative precisions ε_f and ε_g should be judiciously determined. We often use $\varepsilon = \varepsilon_g = \varepsilon_f$. Our stopping criterion in the inner loop reads

$$\|\mathbf{g}^P(\mathbf{y}^i, \boldsymbol{\lambda}^k, \rho)\| \le \min\{M_k\|\mathsf{B}\mathbf{y}^i - \mathbf{c}\|, \eta\} \ \text{ or } \ (9.57),$$

so that the inner loop is interrupted when either the solution or a new iterate $\mathbf{x}^k = \mathbf{y}^i$ is found.

The parameter η is used to define the initial bound on the feasibility error which is used to control the update of M. The algorithm does not seem to be sensitive with respect to η; we use $\eta = 0.1\|\mathbf{b}\|$.

The parameter β is used to increase the precision control. Our experience indicates that $\beta = 0.2$ is a reasonable choice.

The regularization parameter ρ should compromise the fast speed of the outer loop with large ρ and a slow convergence of the algorithm in the inner loop. For the problems arising from the dual formulation of the conditions of equilibrium of contact problems, the choice $\rho \approx \|\mathsf{A}\|$ does not increase the upper bound on the spectrum of the Hessian of the augmented Lagrangian and seems to balance reasonably the speed of the inner and outer iterations.

The basic strategy for initialization of M_0 is based on the relation

$$M_k^2 < \rho\lambda_{\min}(\mathsf{A}),$$

which guarantees sufficient increase of the augmented Lagrangian. We use

$$M_0^2 \approx 100\rho\lambda_{\min}(\mathsf{A}),$$

which allows fast early updates of the Lagrange multipliers. Notice that M_k is adjusted automatically, so the performance of the algorithm is not sensitive to M_0.

If the Hessian $\mathsf{H} = \mathsf{A} + \rho\mathsf{B}\mathsf{B}^T$ of L is ill-conditioned and there is an approximation M of H that can be used as preconditioner, then we can use the preconditioning strategies introduced in Sect. 8.6. We report effective problem dependent preconditioners for contact problems in Chap. 16.

9.15 Comments and References

This chapter is based on the research the starting point of which was the algorithm introduced by Conn, Gould, and Toint [12]; they adapted the augmented Lagrangian method of Powell [2] and Hestenes [1] to the solution of problems with a nonlinear cost function subject to nonlinear equality constraints and bound constraints. Conn, Gould, and Toint worked with the increasing penalty parameter. They also proved that the potentially troublesome penalty parameter ρ_k is bounded and the algorithm converges to a solution also with asymptotically exact solutions of auxiliary problems [12]. Moreover, they used their algorithm to develop the LANCELOT [5] package for the solution of more general nonlinear optimization problems. More references can be found in their comprehensive book on trust region methods [13]. An excellent reference for the application of augmented Lagrangian for more general problems can be found in Birgin and Martínez [14]. For the augmented Lagrangian method with filter, see Friedlander and Leyfer [15].

The SMALSE and SMALBE algorithms differ from the original algorithm in two points. The first one is the adaptive precision control introduced for bound and equality constrained problems by Dostál, Friedlander, and Santos [7]. These authors also proved the basic convergence results for the problems with a regular solution, including the linear convergence of both the Lagrange multipliers and the feasibility

error for a large initial penalty parameter ρ_0. The algorithm presented in [7] generated a forcing sequence for the increase of penalty parameter which was sufficient to get the convergence results.

The next modification, the update rule for the parameter ρ_k of the original SMALBE which enforced a sufficient monotonic increase of $L(\mathbf{x}^k, \mu^k, \rho_k)$, was first published by Dostál [16]. The convergence analysis included the optimality of the outer loop and the bound on the penalty parameter.

The possibility to keep the regularization parameter fixed was first mentioned in the book [11]. The algorithm for the bound and equality constraints which keeps the regularization fixed was coined SMALBE-M. This innovation was important for experimental demonstration of numerical scalability of the algorithms for multibody contact problems. The convergence analysis shows that it is possible to combine both strategies.

The first optimality results for the bound and equality constrained problems were proved by Dostál and Horák for the penalty method [17, 18]. The optimality of SMALBE with the auxiliary problems solved by MPRGP was proved in Dostál [19]; the generalization of the results achieved earlier for the penalty method was based on a well-known observation that the basic augmented Lagrangian algorithm can be considered as a variant of the penalty method (see, e.g., Bertsekas [20, Sect. 4.4]). The generalization to the solution of problems with more general separable constraints was easy after the development of the algorithms discussed in Chap. 7. The presentation of SMALSE-M given here is based on our earlier work, Dostál and Kučera [8], and Dostál and Kozubek [9]. The SMALSE-Mw algorithm adapted for the elliptic constraints with strong excentricity appeared in Bouchala et al. [21]. Linear convergence for SMALBE-M has been proved in Dostál et al. [10]. If applied to the bound and equality constrained problem (9.42), all these algorithms generate identical iterates. Effective heuristic modifications can be found in Hapla [22].

References

1. Hestenes, C.: Multiplier and gradient methods. J. Optim. Theory Appl. **4**, 303–320 (1969)
2. Powell, M.J.D.: A Method for Nonlinear Constraints in Minimization Problems. Optimization, pp. 283–298. Academic Press, New York (1969)
3. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Methods. Academic Press, London (1982)
4. Glowinski, R., Le Tallec, P.: Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics. SIAM, Philadelphia (1989)
5. Conn, A.R., Gould, N.I.M., Toint, PhL: LANCELOT: A FORTRAN Package for Large Scale Nonlinear Optimization (Release A). Springer Series in Computational Mathematics, vol. 17. Springer, New York (1992)
6. Hager, W.W.: Analysis and implementation of a dual algorithm for constraint optimization. J. Optim. Theory Appl. **79**, 37–71 (1993)
7. Dostál, Z., Friedlander, A., Santos, S.A.: Augmented Lagrangians with adaptive precision control for quadratic programming with simple bounds and equality constraints. SIAM J. Optim. **13**, 1120–1140 (2003)

8. Dostál, Z., Kučera, R.: An optimal algorithm for minimization of quadratic functions with bounded spectrum subject to separable convex inequality and linear equality constraints. SIAM J. Optim. **20**(6), 2913–2938 (2010)
9. Dostál, Z., Kozubek, T.: An optimal algorithm with superrelaxation for minimization of a quadratic function subject to separable constraints with applications. Math. Program. Ser. A **135**, 195–220 (2012)
10. Dostál, Z., Brzobohatý, T., Horák, D., Kozubek, T., Vodstrčil, P.: On R-linear convergence of semi-monotonic inexact augmented Lagrangians for bound and equality constrained quadratic programming problems with application. Comput. Math. Appl. **67**(3), 515–526 (2014)
11. Dostál, Z.: Optimal Quadratic Programming Algorithms, with Applications to Variational Inequalities, 1st edn. Springer, New York (2009)
12. Conn, A.R., Gould, N.I.M., Toint, Ph.L.: A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. SIAM J. Numer. Anal. **28**, 545–572 (1991)
13. Conn, A.R., Gould, N.I.M., Toint, Ph.L.: Trust Region Methods. SIAM, Philadelphia (2000)
14. Birgin, E.M., Martínez, J.M.: Practical Augmented Lagrangian Method. SIAM, Philadelphia (2014)
15. Friedlander, M.P., Leyfer, S.: Global and finite termination of a two-phase augmented Lagrangian filter method for general quadratic programs. SIAM J. Sci. Comput. **30**(4), 1706–1729 (2008)
16. Dostál, Z.: Inexact semi-monotonic augmented Lagrangians with optimal feasibility convergence for quadratic programming with simple bounds and equality constraints. SIAM J. Numer. Anal. **43**(1), 96–115 (2005)
17. Dostál, Z., Horák, D.: Scalable FETI with optimal dual penalty for a variational inequality. Numer. Linear Algebra Appl. **11**(5–6), 455–472 (2004)
18. Dostál, Z., Horák, D.: Scalable FETI with optimal dual penalty for semicoercive variational inequalities. Contemp. Math. **329**, 79–88 (2003)
19. Dostál, Z.: An optimal algorithm for bound and equality constrained quadratic programming problems with bounded spectrum. Computing **78**, 311–328 (2006)
20. Bertsekas, D.P.: Nonlinear Optimization. Athena Scientific, Belmont (1999)
21. Bouchala, J., Dostál, Z., Kozubek, T., Pospíšil, L., Vodstrčil, P.: On the solution of convex QPQC problems with elliptic and other separable constraints. Appl. Math. Comput. **247**(15), 848–864 (2014)
22. Hapla, V.: Massively parallel quadratic programming solvers with applications in mechanics. Ph.D. thesis, FEECS VSB–Technical University of Ostrava, Ostrava (2016)

Part III
Scalable Algorithms for Contact Problems

Chapter 10
TFETI for Scalar Problems

We shall first illustrate the ideas of scalable domain decomposition algorithms for contact problems by describing the solution of two scalar problems governed by elliptic boundary variational inequalities. The problems proposed by Ivan Hlaváček deal with the equilibrium of two membranes with prescribed unilateral conditions on the parts of their boundaries and have a structure similar to multibody contact problems of elasticity. The scalar variational inequalities are of independent interest as they describe the steady state solutions of problems arising in various fields of mathematical physics (see, e.g., Duvaut and Lions [1]).

Our presentation is based on the *FETI* (Finite Element Tearing and Interconnecting) method, which was proposed as a parallel solver for linear problems arising from the discretization of elliptic partial differential equations. The basic idea is to decompose the domain into nonoverlapping subdomains that are "glued" by equality constraints. Using the duality, the original problem is reduced to a small, relatively well-conditioned QP problem in Lagrange multipliers.

Here we introduce a variant of the FETI method called *Total FETI (TFETI)*, which differs from original FETI in a way which is used to implement Dirichlet boundary conditions. While FETI assumes that the subdomains inherit their Dirichlet boundary conditions from the original problem, TFETI enforces them by Lagrange multipliers. Such approach simplifies the implementation as the stiffness matrices of "floating" subdomains have a priori known kernels that define the coarse problem which does not depend on prescribed displacements. If the procedure is combined with the preconditioning by the "natural coarse grid" of rigid body motions, the regular condition number of the Hessian matrix of the dual energy function becomes uniformly bounded.

Though the FETI methods are well established as efficient tools for the parallel solution of linear problems, they are even more efficient for the solution of boundary variational inequalities. The reasons are that the duality reduces the inequality constraints to bound constraints and the "natural coarse grid" defines a sufficiently small subspace with a solution. As a result, we get a small convex QP problem with bound and equality constraints and a well-conditioned Hessian that can be solved by the SMALBE-M and MPRGP algorithms with asymptotically linear complexity.

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3_10

10.1 Two Membranes in Unilateral Contact

We shall reduce our analysis to two scalar model problems. Let $\Omega = \Omega^1 \cup \Omega^2$, $\Omega^1 = (0,1) \times (0,1)$, and $\Omega^2 = (1,2) \times (0,1)$ denote open domains with the boundaries Γ^1, Γ^2. Let the parts Γ_U^i, Γ_F^i, and

$$\Gamma_C^i = \Gamma_C = \{(1, y) \in \mathbb{R}^2 : y \in (0,1)\}$$

of Γ^i be formed by the sides of Ω^i, $i = 1, 2$. Let $f : \Omega \to \mathbb{R}$ denote a given continuous function. Our goal is to find a sufficiently smooth (u^1, u^2) satisfying

$$-\Delta u^i = f \quad \text{in } \Omega^i, \quad u^i = 0 \quad \text{on } \Gamma_U^i, \quad \frac{\partial u^i}{\partial \mathbf{n}^i} = 0 \quad \text{on } \Gamma_F^i, \quad i = 1, 2, \tag{10.1}$$

where $\mathbf{n}^i$ denotes the outer unit normal, together with the conditions given on $\Gamma_C = \Gamma_C^1 = \Gamma_C^2$

$$u^2 - u^1 \geq 0, \quad \frac{\partial u^2}{\partial \mathbf{n}^2} \geq 0, \quad \frac{\partial u^2}{\partial \mathbf{n}^2}(u^2 - u^1) = 0, \quad \frac{\partial u^1}{\partial \mathbf{n}^1} + \frac{\partial u^2}{\partial \mathbf{n}^2} = 0. \tag{10.2}$$

The solution u can be interpreted as the displacement of two membranes that are fixed on Γ_U, pressed vertically by the traction of the density f, and pulled horizontally by the unit density traction along Γ_F and the part of Γ_C where $u^1 < u^2$. Moreover, the left edge of the right membrane is not allowed to penetrate below the right edge of the left membrane and the latter can only be pressed down.

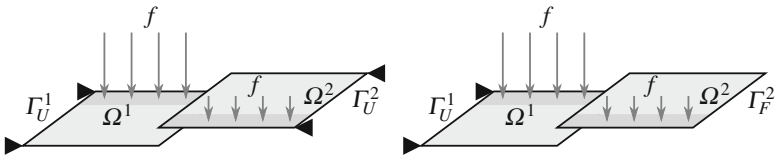

Fig. 10.1 Coercive (*left*) and semicoercive (*right*) model problems

We shall distinguish two cases. In the first case, both membranes are fixed on the outer edges as in Fig. 10.1 left, so that

$$\Gamma_U^1 = \{(0, y) \in \mathbb{R}^2 : y \in [0,1]\}, \quad \Gamma_U^2 = \{(2, y) \in \mathbb{R}^2 : y \in [0,1]\}.$$

Since the Dirichlet conditions are prescribed on the parts $\Gamma_U^i,\ i = 1, 2$, of the boundaries with a positive measure, it follows that a solution exists and is necessarily unique [2, 3]. In the second case, only the left membrane is fixed on the outer edge and the right membrane has no prescribed vertical displacement as in Fig. 10.1 right, so that

$$\Gamma_U^1 = \{(0, y) \in \mathbb{R}^2 : y \in [0, 1]\}, \quad \Gamma_U^2 = \emptyset.$$

To guarantee the solvability and uniqueness, we shall assume

$$\int_{\Omega^2} f \, \mathrm{d}\Omega < 0.$$

More details about this model problem may be found, e.g., in [4].

10.2 Variational Formulation

To reduce the requirements on the smoothness of data and a solution $u = (u_1, u_2)$ of (10.1) and (10.2), let us reformulate the problem in a variational form, which requires that the relations are satisfied rather in average than point-wise. This approach opens a convenient way to the formulation of the results concerning the existence and uniqueness of a solution and to the application of efficient QP solvers to the solution of discretized problems. The variational form of (10.1) and (10.2) uses the forms defined on suitable Sobolev spaces.

Let $H^1(\Omega^i),\ i = 1, 2$, denote the Sobolev spaces of the first order in the space $L^2(\Omega^i)$ of the functions defined on Ω^i the squares of which are integrable in the sense of Lebesgue. Let

$$V^i = \left\{v^i \in H^1(\Omega^i) : v^i = 0 \quad \text{on} \quad \Gamma_U^i\right\}$$

denote the closed subspaces of $H^1(\Omega^i),\ i = 1, 2$, and let

$$V = V^1 \times V^2 \qquad \text{and} \qquad \mathscr{K} = \left\{(v^1, v^2) \in V : v^2 - v^1 \geq 0 \quad \text{on} \quad \Gamma_C\right\}$$

denote the closed subspace and the closed convex subset of

$$H = H^1(\Omega^1) \times H^1(\Omega^2),$$

respectively. The relations on the boundaries are in terms of traces. On H we shall consider the $L^2(\Omega)$ scalar product

$$(u, v) = \sum_{i=1}^{2} \int_{\Omega^i} u^i v^i \, \mathrm{d}\Omega,$$

the symmetric bilinear form

$$a(u,v) = \sum_{i=1}^{2} \int_{\Omega^i} \left(\frac{\partial u^i}{\partial x_1} \frac{\partial v^i}{\partial x_1} + \frac{\partial u^i}{\partial x_2} \frac{\partial v^i}{\partial x_2} \right) \mathrm{d}\Omega,$$

and the linear form

$$\ell(v) = (f,v) = \sum_{i=1}^{2} \int_{\Omega^i} f^i v^i \, \mathrm{d}\Omega, \quad f^i = f|\Omega^i.$$

To get variational conditions that are satisfied by each sufficiently smooth solution $u = (u^1, u^2)$ of (10.1) and (10.2), let $v = (v^1, v^2) \in C^1(\Omega^1) \times C^1(\Omega^2) \cap \mathscr{K}$. Using the Green Theorem 4.2, the definition of $\mathscr{K}$, and (10.1) and (10.2), we get

$$-(\Delta u, v) = (\nabla u, \nabla v) - \int_{\Gamma_C} \frac{\partial u^1}{\partial \mathbf{n}^1} v_1 \, \mathrm{d}\Gamma - \int_{\Gamma_C} \frac{\partial u^2}{\partial \mathbf{n}^2} v_2 \, \mathrm{d}\Gamma \tag{10.3}$$

$$= a(u,v) + \int_{\Gamma_C} \frac{\partial u^2}{\partial \mathbf{n}^2} (v^1 - v^2) \, \mathrm{d}\Gamma \le a(u,v). \tag{10.4}$$

For $\mathbf{x} \in \Omega^i$, we define $u(\mathbf{x}) = u^i(\mathbf{x})$. Since for any $v \in C^1(\Omega^1) \times C^1(\Omega^2)$

$$-(\Delta u, v) = (f,v)$$

and a solution u satisfies

$$\int_{\Gamma_C} \frac{\partial u^2}{\partial \mathbf{n}^2} (u^1 - u^2) \, \mathrm{d}\Gamma = 0,$$

we can use the assumption $v - u \in C^1(\Omega^1) \times C^1(\Omega^2)$ to get

$$a(u, v-u) - \ell(v-u) = -(\Delta u, v-u) + \int_{\Gamma_C} \frac{\partial u^2}{\partial \mathbf{n}^2} (v^2 - v^1) \, \mathrm{d}\Gamma - (f, v-u)$$

$$= \int_{\Gamma_C} \frac{\partial u^2}{\partial \mathbf{n}^2} (v^2 - v^1) \, \mathrm{d}\Gamma \ge 0. \tag{10.5}$$

The latter condition, i.e.,

$$a(u, v-u) \ge \ell(v-u), \quad v \in \mathscr{K} \cap C^1(\Omega^1) \times C^1(\Omega^2),$$

is by Theorem 4.5 just the condition which satisfies the minimizer of

$$q(v) = \frac{1}{2} a(v,v) - \ell(v) \ \text{ subject to } \ v \in \mathscr{K} \cap C^1(\Omega^1) \times C^1(\Omega^2). \tag{10.6}$$

The classical formulation of the conditions of equilibrium does not describe realistic situations, e.g., if f is discontinuous, then the description of equilibrium by (10.1) and (10.2) is not complete, but the problem to find $u \in \mathcal{K}$ such that

$$q(u) \leq q(v), \quad v \in \mathcal{K}, \tag{10.7}$$

is defined for any $f \in L^2(\Omega^1) \times L^2(\Omega^2)$. Using the arguments based on the coercivity of q, it is possible to prove that (10.7) has a solution [2] and that any sufficiently smooth solution of (10.7) solves (10.1) and (10.2). Moreover, if f is piece-wise discontinuous, then it can be shown that the solution of variational problem (10.7) satisfies additional conditions of equilibrium [2].

10.3 Tearing and Interconnecting

So far, we have used only the natural decomposition of the spatial domain Ω into Ω^1 and Ω^2. However, to enable efficient application of domain decomposition methods, we can optionally decompose each Ω^i into $p = 1/H \times 1/H$ square subdomains $\Omega^{i1}, \ldots, \Omega^{ip}$ as in Fig. 10.2. We shall call H a *decomposition parameter*.

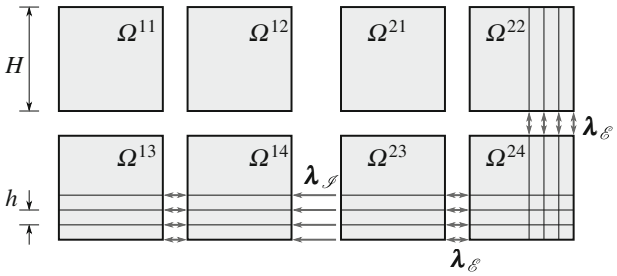

Fig. 10.2 Domain decomposition and discretization

The continuity of a global solution in Ω^1 and Ω^2 can be enforced by the "gluing" conditions

$$u^{ij}(\mathbf{x}) = u^{ik}(\mathbf{x}), \tag{10.8}$$

$$\nabla u^{ij} \cdot \mathbf{n}^{ij} = -\nabla u^{ik} \cdot \mathbf{n}^{ik}, \tag{10.9}$$

which should be satisfied by the traces of u^{ij} and u^{ik} on $\Gamma^{ij,ik} = \Gamma^{ij} \cap \Gamma^{ik}$.

To get a variational formulation of the decomposed problem, let

$$V^{ij} = \left\{ v^{ij} \in H^1(\Omega^{ij}) : v^{ij} = 0 \quad \text{on} \quad \Gamma_U \cap \Gamma^{ij} \right\}, \quad i = 1, 2, \quad j = 1, \ldots p,$$

denote the closed subspaces of $H^1(\Omega^{ij})$ and let

$$V_{DD} = (V^{11} \times \cdots \times V^{1p}) \times (V^{21} \times \cdots \times V^{2p}),$$
$$\mathscr{K}_{DD} = \left\{ v \in V_{DD} : v^{2j} - v^{1i} \geq 0 \text{ on } \Gamma_C^{1i} \cap \Gamma_C^{2j} \text{ and } v^{ij} = v^{ik} \text{ on } \Gamma^{ij,ik} \right\}.$$

The relations on the boundaries are again in terms of traces. On V_{DD}, we shall define the scalar product

$$(u, v) = \sum_{i=1}^{2} \sum_{j=1}^{p} \int_{\Omega^{ij}} u^{ij} v^{ij} \, \mathrm{d}\Omega,$$

the symmetric bilinear form

$$a(u, v) = \sum_{i=1}^{2} \sum_{j=1}^{p} \int_{\Omega^{ij}} \left(\frac{\partial u^{ij}}{\partial x_1} \frac{\partial v^{ij}}{\partial x_1} + \frac{\partial u^{ij}}{\partial x_2} \frac{\partial v^{ij}}{\partial x_2} \right) \mathrm{d}\Omega,$$

and the linear form

$$\ell(v) = (f, v) = \sum_{i=1}^{2} \sum_{j=1}^{p} \int_{\Omega^{ij}} f^{ij} v^{ij} \, \mathrm{d}\Omega,$$

where $f^{ij} \in L^2(\Omega^{ij})$ denotes the restriction of f to Ω^{ij}.

Observing that for any $v \in \mathscr{K}_{DD}$

$$\int_{\Gamma_G^{ij,ik}} \frac{\partial u^{ij}}{\partial \mathbf{n}^{ij}} v^{ij} \, \mathrm{d}\Gamma + \int_{\Gamma_G^{ij,ik}} \frac{\partial u^{ik}}{\partial \mathbf{n}^{ik}} v^{ik} \, \mathrm{d}\Gamma = \int_{\Gamma_G^{ij,ik}} \left(\frac{\partial u^{ij}}{\partial \mathbf{n}^{ij}} + \frac{\partial u^{ik}}{\partial \mathbf{n}^{ik}} \right) v^{ij} \, \mathrm{d}\Gamma = 0,$$

we get that relations (10.3)–(10.5) remain valid also for the decomposed problem and $u \in \mathscr{K}_{DD}$ solves

$$a(u, v - u) \geq \ell(v - u), \quad v \in \mathscr{K}_{DD}. \tag{10.10}$$

In what follows, we shall consider the equivalent problem $u \in \mathscr{K}_{DD}$ such that

$$q(u) \leq q(v), \quad q(v) = \frac{1}{2} a(v, v) - \ell(v), \quad v \in \mathscr{K}_{DD}. \tag{10.11}$$

10.4 Discretization

After introducing regular grids with the discretization parameter h in the subdomains Ω^{ij}, so that they match across the interfaces $\Gamma^{ij,kl}$, indexing contiguously the nodes and entries of corresponding vectors in the subdomains, and using a Lagrangian

finite element discretization, we get the discretized version of problem (10.11) with auxiliary domain decomposition

$$\min \frac{1}{2}\mathbf{u}^T \mathsf{K}\mathbf{u} - \mathbf{f}^T\mathbf{u} \quad \text{s.t.} \quad \mathsf{B}_I\mathbf{u} \le \mathbf{o} \quad \text{and} \quad \mathsf{B}_E\mathbf{u} = \mathbf{o}. \tag{10.12}$$

In (10.12), $\mathsf{K} \in \mathbb{R}^{n\times n}$ denotes a block diagonal SPS stiffness matrix, the full rank matrices B_I and B_E describe the discretized non-penetration and gluing conditions, respectively, and $\mathbf{f}$ represents the discrete analog of the linear term $\ell(u)$. If we replace the couples of indices by single indices, we can write the stiffness matrix and the vectors in the block form

$$\mathsf{K} = \begin{bmatrix} \mathsf{K}_1 & \mathsf{O} & \dots & \mathsf{O} \\ \mathsf{O} & \mathsf{K}_2 & \dots & \mathsf{O} \\ \dots & \dots & \dots & \dots \\ \mathsf{O} & \mathsf{O} & \dots & \mathsf{K}_s \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \dots \\ \mathbf{u}_s \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} \mathbf{f}_1 \\ \dots \\ \mathbf{f}_s \end{bmatrix}, \quad s = 2p.$$

The rows of B_E and B_I are filled with zeros except 1 and -1 in the positions that correspond to the nodes with the same coordinates on the artificial or contact boundaries, respectively. We get three types of equality constraints as in Fig. 10.3.

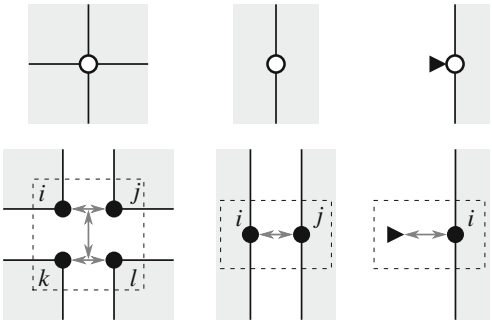

Fig. 10.3 Three types of constraints

If $\mathbf{b}_i$ denotes a row of B_I or B_E, then $\mathbf{b}_i$ does not have more than four nonzero entries. The continuity of the solution in the "wire basket" (see Fig. 10.3 left) and on the interface (see Fig. 10.3 middle) or the fulfillment of Dirichlet's boundary conditions (see Fig. 10.3 right) are enforced by the equalities

$$u_i = u_j,\ u_k = u_\ell,\ u_i + u_j = u_k + u_l; \quad u_i = u_j; \quad u_i = 0;$$

respectively, which can be expressed by means of the vectors

$$\mathbf{b}_{ij} = (\mathbf{s}_i - \mathbf{s}_j)^T,\ \mathbf{b}_{k\ell} = (\mathbf{s}_k - \mathbf{s}_\ell)^T,\ \mathbf{b}_{ijkl} = (\mathbf{s}_i + \mathbf{s}_j - \mathbf{s}_k - \mathbf{s}_l)^T;$$
$$\mathbf{b}_{ij} = (\mathbf{s}_i - \mathbf{s}_j)^T;\ \mathbf{b}_i = \mathbf{s}_i^T;$$

where $\mathbf{s}_i$ denotes the ith column of the identity matrix I_n. The continuity of the solution across the subdomains interface (see Fig. 10.3 middle) is implemented by

$$\mathbf{b}_{ij}\mathbf{x} = 0,$$

so that $\mathbf{b}_{ij}\mathbf{x}$ denotes the jump across the boundary.

The non-penetration is enforced similarly. If i and j are the indices of matching nodes on Γ_C^1 and Γ_C^2, respectively, then any feasible nodal displacements satisfy

$$\mathbf{b}_{ij}\mathbf{x} \leq 0.$$

The construction of the matrices B_E and B_I guarantees that any couple of their rows is orthogonal. We can easily achieve by scaling that $\mathsf{B} = [\mathsf{B}_E^T, \mathsf{B}_I^T]^T$ satisfies

$$\mathsf{B}\mathsf{B}^T = \mathsf{I}.$$

10.5 Dual Formulation

Our next step is to simplify the problem using the duality theory, in particular we replace the general inequality constraints

$$\mathsf{B}_I\mathbf{u} \leq \mathbf{o}$$

by the nonnegativity constraints. To this end, let us define the Lagrangian associated with problem (10.12) by

$$L(\mathbf{u}, \boldsymbol{\lambda}_I, \boldsymbol{\lambda}_E) = \frac{1}{2}\mathbf{u}^T\mathsf{K}\mathbf{u} - \mathbf{f}^T\mathbf{u} + \boldsymbol{\lambda}_I^T\mathsf{B}_I\mathbf{u} + \boldsymbol{\lambda}_E^T\mathsf{B}_E\mathbf{u}, \tag{10.13}$$

where $\boldsymbol{\lambda}_I$ and $\boldsymbol{\lambda}_E$ are the Lagrange multipliers associated with the inequalities and equalities, respectively. Introducing the notation

$$\boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{\lambda}_I \\ \boldsymbol{\lambda}_E \end{bmatrix} \quad \text{and} \quad \mathsf{B} = \begin{bmatrix} \mathsf{B}_I \\ \mathsf{B}_E \end{bmatrix},$$

we can observe that $\mathsf{B} \in \mathbb{R}^{m\times n}$ is a full rank matrix and write the Lagrangian briefly as

$$L(\mathbf{u}, \boldsymbol{\lambda}) = \frac{1}{2}\mathbf{u}^T\mathsf{K}\mathbf{u} - \mathbf{f}^T\mathbf{u} + \boldsymbol{\lambda}^T\mathsf{B}\mathbf{u}.$$

Thus the solution satisfies the KKT conditions, including

$$\mathsf{K}\mathbf{u} - \mathbf{f} + \mathsf{B}^T\boldsymbol{\lambda} = \mathbf{o}. \tag{10.14}$$

Equation (10.14) has a solution if and only if

$$\mathbf{f} - \mathsf{B}^T\boldsymbol{\lambda} \in \mathrm{Im}\mathsf{K}, \tag{10.15}$$

which can be expressed more conveniently by means of a matrix R the columns of which span the null space of K as

$$\mathsf{R}^T(\mathbf{f} - \mathsf{B}^T\boldsymbol{\lambda}) = \mathbf{o}. \tag{10.16}$$

The matrix R can be formed directly so that each floating subdomain is assigned to a column of R with ones in the positions of the nodal variables that belong to the subdomain and zeros elsewhere. It may be checked that $\mathsf{R}^T\mathsf{B}^T$ is a full rank matrix.

Now assume that $\boldsymbol{\lambda}$ satisfies (10.15), so that we can evaluate $\boldsymbol{\lambda}$ from (10.14) by means of any (left) generalized matrix K^+ which satisfies

$$\mathsf{K}\mathsf{K}^+\mathsf{K} = \mathsf{K}. \tag{10.17}$$

It may be verified directly that if $\mathbf{u}$ solves (10.14), then there is a vector α such that

$$\mathbf{u} = \mathsf{K}^+(\mathbf{f} - \mathsf{B}^T\boldsymbol{\lambda}) + \mathsf{R}\alpha. \tag{10.18}$$

For the effective evaluation of the generalized inverse, we can use

$$\mathsf{K}^\# = \mathrm{diag}(\mathsf{K}_1^\#, \ldots, \mathsf{K}_{2p}^\#),$$

where $\mathsf{K}_i^\#$ is defined in (2.6). Using Lemma 2.1, we can check that we can get a full rank submatrix A_i of K_i, which appears in the definition of $\mathsf{K}_i^\#$, by deleting any row and corresponding column of K_i. The best conditioning of $\mathsf{K}_i^\#$ can be achieved by deleting those corresponding to a node near the center of Ω^i [5]. The action of $\mathsf{K}^\#$ can be evaluated by the Cholesky decomposition. See Sect 11.7 for more details and an alternative procedure.

Using Proposition 3.13, we can find $\boldsymbol{\lambda}$ by solving the minimization problem

$$\min \theta(\boldsymbol{\lambda}) \quad \text{s.t.} \quad \boldsymbol{\lambda}_{\mathscr{I}} \geq \mathbf{o} \quad \text{and} \quad \mathsf{R}^T(\mathbf{f} - \mathsf{B}^T\boldsymbol{\lambda}) = \mathbf{o}, \tag{10.19}$$

where

$$\theta(\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\lambda}^T\mathsf{B}\mathsf{K}^+\mathsf{B}^T\boldsymbol{\lambda} - \boldsymbol{\lambda}^T\mathsf{B}\mathsf{K}^+\mathbf{f}. \tag{10.20}$$

Notice that θ is obtained from the dual function Θ defined by (3.50) by changing the signs and omitting the constant term. Once the solution $\widehat{\boldsymbol{\lambda}}$ of (10.19) is known, the vector $\widehat{\mathbf{u}}$ which solves (10.12) can be evaluated by (10.18) and the formula (3.67). We get

$$\boldsymbol{\alpha} = -(\mathsf{R}^T\widetilde{\mathsf{B}}^T\widetilde{\mathsf{B}}\mathsf{R})^{-1}\mathsf{R}^T\widetilde{\mathsf{B}}^T\widetilde{\mathsf{B}}\mathsf{K}^+(\mathbf{f} - \mathsf{B}^T\widehat{\boldsymbol{\lambda}}), \tag{10.21}$$

where $\widetilde{\mathsf{B}} = [\widetilde{\mathsf{B}}_I^T, \mathsf{B}_E^T]^T$, and the matrix $\widetilde{\mathsf{B}}_I$ is formed by the rows $\mathbf{b}_i$ of B_I that correspond to the positive components of the solution $\widehat{\boldsymbol{\lambda}}_I$ characterized by $\widehat{\lambda}_i > 0$.

10.6 Natural Coarse Grid

Even though problem (10.19) is much more suitable for computations than (10.12) and was used to effective solving of discretized variational inequalities [6], further improvement can be achieved using orthogonal projectors associated with the feasible set. Let us denote

$$\begin{aligned} \mathsf{F} &= \mathsf{B}\mathsf{K}^+\mathsf{B}^T, \quad \widetilde{\mathbf{d}} = \mathsf{B}\mathsf{K}^+\mathbf{f}, \\ \widetilde{\mathsf{G}} &= \mathsf{R}^T\mathsf{B}^T, \quad \widetilde{\mathbf{e}} = \mathsf{R}^T\mathbf{f}, \end{aligned}$$

and let T denote a regular matrix that defines orthonormalization of the rows of $\widetilde{\mathsf{G}}$ so that the matrix

$$\mathsf{G} = \mathsf{T}\widetilde{\mathsf{G}}$$

has orthonormal rows. After denoting

$$\mathbf{e} = \mathsf{T}\widetilde{\mathbf{e}},$$

problem (10.19) reads

$$\min \quad \frac{1}{2}\boldsymbol{\lambda}^T\mathsf{F}\boldsymbol{\lambda} - \boldsymbol{\lambda}^T\widetilde{\mathbf{d}} \quad \text{s.t.} \quad \boldsymbol{\lambda}_I \geq \mathbf{o} \quad \text{and} \quad \mathsf{G}\boldsymbol{\lambda} = \mathbf{e}. \tag{10.22}$$

Next we shall transform the problem of minimization on the subset of the affine space to that on the subset of the vector space by looking for the solution of (10.22) in the form

$$\boldsymbol{\lambda} = \boldsymbol{\mu} + \widetilde{\boldsymbol{\lambda}}, \quad \text{where} \quad \mathsf{G}\widetilde{\boldsymbol{\lambda}} = \mathbf{e}.$$

The following rather trivial lemma shows that we can even find $\widetilde{\boldsymbol{\lambda}}$ such that $\widetilde{\boldsymbol{\lambda}}_I \geq \mathbf{o}$.

Lemma 10.1 *There is* $\widetilde{\boldsymbol{\lambda}}_I \geq 0$ *such that* $\mathsf{G}\widetilde{\boldsymbol{\lambda}} = \widetilde{\mathbf{e}}$.

Proof Take

$$\widetilde{\boldsymbol{\lambda}} = \arg\min \frac{1}{2}\|\boldsymbol{\lambda}\|^2 \quad \text{s.t.} \quad \boldsymbol{\lambda}_I \geq \mathbf{o} \quad \text{and} \quad \mathsf{G}\boldsymbol{\lambda} = \mathbf{e}.$$

□

To carry out the transformation, denote $\boldsymbol{\lambda} = \boldsymbol{\mu} + \widetilde{\boldsymbol{\lambda}}$, so that

$$\frac{1}{2}\boldsymbol{\lambda}^T \mathsf{F}\boldsymbol{\lambda} - \boldsymbol{\lambda}^T \widetilde{\mathbf{d}} = \frac{1}{2}\boldsymbol{\mu}^T \mathsf{F}\boldsymbol{\mu} - \boldsymbol{\mu}^T(\widetilde{\mathbf{d}} - \mathsf{F}\widetilde{\boldsymbol{\lambda}}) + \frac{1}{2}\widetilde{\boldsymbol{\lambda}}^T \mathsf{F}\widetilde{\boldsymbol{\lambda}} - \widetilde{\boldsymbol{\lambda}}^T \widetilde{\mathbf{d}}$$

and problem (10.22) is, after returning to the old notation, equivalent to

$$\min \quad \frac{1}{2}\boldsymbol{\lambda}^T \mathsf{F}\boldsymbol{\lambda} - \boldsymbol{\lambda}^T \mathbf{d} \quad \text{s.t.} \quad \mathsf{G}\boldsymbol{\lambda} = \mathbf{o} \quad \text{and} \quad \boldsymbol{\lambda}_I \geq -\widetilde{\boldsymbol{\lambda}}_I \tag{10.23}$$

with $\mathbf{d} = \widetilde{\mathbf{d}} - \mathsf{F}\widetilde{\boldsymbol{\lambda}}$ and $\widetilde{\boldsymbol{\lambda}}_I \geq \mathbf{o}$.

Our final step is based on the observation that problem (10.23) is equivalent to

$$\min \quad \overline{\theta}_\rho(\boldsymbol{\lambda}) \quad \text{s.t.} \quad \mathsf{G}\boldsymbol{\lambda} = \mathbf{o} \quad \text{and} \quad \lambda_I \geq -\widetilde{\boldsymbol{\lambda}}_I, \tag{10.24}$$

where ρ is a positive constant and

$$\overline{\theta}_\rho(\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\lambda}^T \mathsf{H}\boldsymbol{\lambda} - \boldsymbol{\lambda}^T \mathsf{P}\mathbf{d}, \quad \mathsf{H} = \mathsf{PFP} + \rho\mathsf{Q}, \quad \mathsf{Q} = \mathsf{G}^T\mathsf{G}, \quad \mathsf{P} = \mathsf{I} - \mathsf{Q}. \tag{10.25}$$

The matrices P and Q are the orthogonal projectors on the kernel of G and the image space of G^T, respectively. The regularization term is introduced in order to enable the reference to the results on strictly convex QP problems.

10.7 Bounds on the Spectrum

We shall solve the bound and equality constrained QP problem (10.24) by SMALBE (Algorithm 9.2) with the inner loop implemented by MPRGP (Algorithm 8.2). These algorithms can solve the class of problems (10.24) arising from the discretization of (10.11) with varying discretization and decomposition parameters in a uniformly bounded number of iterations provided there are positive bounds on the spectrum of the Hessian H of the cost function $\overline{\theta}_\rho$.

First observe that $\mathrm{Im}\mathsf{P}$ and $\mathrm{Im}\mathsf{Q}$ are invariant subspaces of H, $\mathrm{Im}\mathsf{P}+\mathrm{Im}\mathsf{Q} = \mathbb{R}^m$ as $\mathsf{P} + \mathsf{Q} = \mathsf{I}$, and for any $\boldsymbol{\lambda} \in \mathbb{R}^m$

$$\mathsf{HP}\boldsymbol{\lambda} = (\mathsf{PFP} + \rho\mathsf{Q})\mathsf{P}\boldsymbol{\lambda} = \mathsf{P}(\mathsf{FP}\boldsymbol{\lambda}) \quad \text{and} \quad \mathsf{HQ}\boldsymbol{\lambda} = (\mathsf{PFP} + \rho\mathsf{Q})\mathsf{Q}\boldsymbol{\lambda} = \rho\mathsf{Q}\boldsymbol{\lambda}.$$

It follows that

$$\sigma(\mathsf{H}|\mathrm{Im}\mathsf{Q}) = \{\rho\},$$

so it remains to find the bounds on

$$\sigma(\mathsf{H}|\mathrm{Im}\mathsf{P}) = \sigma(\mathsf{F}|\mathrm{Im}\mathsf{P}).$$

The next lemma reduces the problem to the analysis of local Schur complements.

Lemma 10.2 *Let there be constants* $0 < c < C$ *such that for each* $\boldsymbol{\lambda} \in \mathbb{R}^m$

$$c\|\boldsymbol{\lambda}\|^2 \le \|\mathsf{B}^T\boldsymbol{\lambda}\|^2 \le C\|\boldsymbol{\lambda}\|^2. \tag{10.26}$$

Then for each $\boldsymbol{\lambda} \in \mathrm{Im}\mathsf{P}$

$$c\left(\max_{i=1,\dots,s}\|\mathsf{S}_i\|\right)^{-1}\|\boldsymbol{\lambda}\|^2 \le \boldsymbol{\lambda}^T\mathsf{F}\boldsymbol{\lambda} \le C\left(\min_{i=1,\dots,s}\overline{\lambda}_{\min}(\mathsf{S}_i)\right)^{-1}\|\boldsymbol{\lambda}\|^2, \tag{10.27}$$

where S_i *denotes the Schur complement of* K_i *with respect to the indices of the interior nodes of* Ω^i *and* $\overline{\lambda}_{\min}(\mathsf{S}_i)$ *denotes the smallest nonzero eigenvalue of* S_i.

Proof Let B be a block matrix which complies with the structure of K, so that

$$\mathsf{F} = \mathsf{B}\mathsf{K}^+\mathsf{B}^T = \left[\mathsf{B}_1, \mathsf{B}_2, \dots, \mathsf{B}_s\right]\begin{bmatrix} \mathsf{K}_1^+ & \mathsf{O} & \dots & \mathsf{O} \\ \mathsf{O} & \mathsf{K}_2^+ & \dots & \mathsf{O} \\ \dots & \dots & \dots & \dots \\ \mathsf{O} & \mathsf{O} & \dots & \mathsf{K}_s^+ \end{bmatrix}\begin{bmatrix} \mathsf{B}_1^T \\ \mathsf{B}_2^T \\ \dots \\ \mathsf{B}_s^T \end{bmatrix} = \sum_{i=1}^{s}\mathsf{B}_i\mathsf{K}_i^+\mathsf{B}_i^T.$$

Since the columns of B_i that correspond to the interior nodes of Ω^i are formed by zero vectors, we can renumber the variables to get $\mathsf{B}_i = [\mathsf{C}_i \ \ \mathsf{O}]$ and

$$c\|\boldsymbol{\lambda}\|^2 \le \sum_{i=1}^{s}\|\mathsf{C}_i^T\boldsymbol{\lambda}\|^2 \le C\|\boldsymbol{\lambda}\|^2.$$

Let us now denote by S_i the Schur complement of K_i, $i = 1, \dots, s$, with respect to the interior variables. Notice that it is well defined, as eliminating of the interior variables of any subdomain amounts to solving the Dirichlet problem which has a unique solution. Moreover, if we denote by $\mathscr{B}$ the set of all indices which correspond to the variables on the boundary of Ω_i and choose a generalized inverse S_i^+, we can use Lemma 2.2 to get

$$\mathsf{F} = \mathsf{B}\mathsf{K}^+\mathsf{B}^T = \sum_{i=1}^{s}\mathsf{B}_i\mathsf{K}_i^+\mathsf{B}_i^T = \sum_{i=1}^{s}\mathsf{C}_i\mathsf{S}_i^+\mathsf{C}_i^T = \mathsf{B}_{*\mathscr{B}}\mathsf{S}^+\mathsf{B}_{*\mathscr{B}}^T,$$

where

$$\mathsf{S} = \mathrm{diag}(\mathsf{S}_1, \dots, \mathsf{S}_s).$$

For the analysis, we shall choose the Moore–Penrose generalized inverse $\mathsf{K}^\dagger$. Observing that for each $\boldsymbol{\lambda} \in \mathbb{R}^m$, $\mathsf{B}^T\mathsf{P}\boldsymbol{\lambda} \in \mathrm{Im}\mathsf{K}$, $\mathrm{Im}\mathsf{K} = \mathrm{Im}\mathsf{K}^\dagger$, and by Lemma 2.2 $\mathsf{B}_{*\mathscr{B}}^T\mathsf{P}\boldsymbol{\lambda} \in \mathrm{Im}\mathsf{S}^\dagger$, we get

$$\boldsymbol{\lambda}^T\mathsf{P}\mathsf{F}\mathsf{P}\boldsymbol{\lambda} = \boldsymbol{\lambda}^T\mathsf{P}\mathsf{B}\mathsf{K}^\dagger\mathsf{B}^T\mathsf{P}\boldsymbol{\lambda} = \boldsymbol{\lambda}^T\mathsf{P}\mathsf{B}_{*\mathscr{B}}\mathsf{S}^\dagger\mathsf{B}_{*\mathscr{B}}^T\mathsf{P}\boldsymbol{\lambda}.$$

Using the assumptions, we get for any $\boldsymbol{\lambda} \in \mathrm{Im}\mathsf{P}$

$$c\lambda_{\max}^{-1}(\mathsf{S})\|\boldsymbol{\lambda}\|^2 \le \lambda_{\max}^{-1}(\mathsf{S})\|\mathsf{B}^T\boldsymbol{\lambda}\|^2 \le \boldsymbol{\lambda}^T\mathsf{B}_{*\mathscr{B}}\mathsf{S}^{\dagger}\mathsf{B}_{*\mathscr{B}}^T\boldsymbol{\lambda} \le \overline{\lambda}_{\min}^{-1}(\mathsf{S})\|\mathsf{B}^T\boldsymbol{\lambda}\|^2 \le C\overline{\lambda}_{\min}^{-1}(\mathsf{S})\|\boldsymbol{\lambda}\|^2.$$

To finish the proof, notice that

$$\lambda_{\max}(\mathsf{S}) = \max_{i=1,\ldots,s} \lambda_{\max}(\mathsf{S}_i) = \max_{i=1,\ldots,s} \|\mathsf{S}_i\| \quad \text{and} \quad \overline{\lambda}_{\min}(\mathsf{S}) = \min_{i=1,\ldots,s} \overline{\lambda}_{\min}(\mathsf{S}_i). \ \square$$

Remark 10.1 Lemma 10.2 indicates that the conditioning of H can be improved by the orthonormalization of the rows of the constraint matrix B.

We have reduced the problem to bound the spectrum $\sigma(\mathsf{H})$ of H to the analysis of the spectra $\sigma(\mathsf{S}_i)$ of the Schur complements of the stiffness matrices of the subdomains with respect to their interior. The following lemma is important in the analysis of optimality of the presented algorithms.

Lemma 10.3 *Let H and h denote the decomposition and discretization parameter, respectively, and let $\mathsf{S}_{H,h}$ denote the Schur complement of the stiffness matrix of a subdomain of Ω with respect to its interior.*

Then there are constants c and C independent of h and H such that for each $\boldsymbol{\lambda} \in \mathrm{Im}\mathsf{S}_{H,h}$

$$c\frac{h}{H}\|\boldsymbol{\lambda}\|^2 \le \boldsymbol{\lambda}^T\mathsf{S}_{H,h}\boldsymbol{\lambda} \le C\|\boldsymbol{\lambda}\|^2. \tag{10.28}$$

Proof See, e.g., Bramble, Pasciak, and Schatz [7] or Pechstein [8, Theorem 2.38 and its proof]. The estimate was a key ingredient of the first optimality analysis of FETI for linear problems by Farhat, Mandel, and Roux [9]. $\square$

The following theorem is now an easy corollary of Lemma 10.2 and Lemma 10.3.

Theorem 10.1 *Let $\rho > 0$ and let $\mathsf{H}_{\rho,H,h}$ denote the Hessian of $\overline{\theta}_\rho$ resulting from the decomposition and discretization of problem (10.7) with the parameters H and h.*

Then there are constants c and C independent of h and H such that for each $\boldsymbol{\lambda} \in \mathbb{R}^n$

$$c\|\boldsymbol{\lambda}\|^2 \le \boldsymbol{\lambda}^T\mathsf{H}_{\rho,H,h}\boldsymbol{\lambda} \le C\frac{H}{h}\|\boldsymbol{\lambda}\|^2. \tag{10.29}$$

Proof Substitute (10.28) into Lemma 10.2, take into account the regularization term $\rho\mathsf{Q}$, and notice that (10.26) is satisfied with $c = 1$ and $C = 4$. $\square$

10.8 Optimality

To show that Algorithm 9.2 (SMALBE) with the inner loop implemented by Algorithm 8.2 (MPRGP) is optimal for the solution of a class of problems arising from the varying discretizations of a given variational inequality, let us introduce new notations which comply with that used in the analysis of algorithms in Part II.

Let

$$\mathscr{T} = \{(H,h) \in \mathbb{R}^2 : H \le 1,\ 0 < 2h \le H, \ \text{ and } \ H/h \in \mathbb{N}\}$$

denote the set of indices, where $\mathbb{N}$ denotes the set of all positive integers. Given a constant $C \ge 2$, we shall define a subset $\mathscr{T}_C$ of $\mathscr{T}$ by

$$\mathscr{T}_C = \{(H,h) \in \mathscr{T} :\ H/h \le C\}.$$

For any $t \in \mathscr{T}$ and $\rho > 0$, we define

$$\begin{aligned} \mathsf{A}_t &= \mathsf{PFP} + \rho\mathsf{Q}, \quad & \mathbf{b}_t &= \mathsf{P}\mathbf{d}, \\ \mathsf{B}_t &= \mathsf{G}, & \ell_I^t &= -\widetilde{\boldsymbol{\lambda}}_I \end{aligned}$$

by the vectors and matrices arising from the discretization of (10.11) with the discretization and decomposition parameters H and h, $t = (H,h)$, so that we get a class of problems

$$\min\ f_t(\boldsymbol{\lambda})\ \text{ s.t. }\ \mathsf{B}_t\boldsymbol{\lambda} = \mathbf{o}\ \text{ and }\ \boldsymbol{\lambda}_I \ge \ell_I^t, \quad t \in \mathscr{T}_C, \tag{10.30}$$

with $f_t(\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\lambda}^T\mathsf{A}_t\boldsymbol{\lambda} - \mathbf{b}_t^T\boldsymbol{\lambda}$. Using Lemma 10.1, we can achieve that $\ell_I^t \le \mathbf{o}$, and using $\mathsf{G}\mathsf{G}^T = \mathsf{I}$, we obtain

$$\|\mathsf{B}_t\| \le 1. \tag{10.31}$$

It follows by Theorem 10.1 that for any $C \ge 2$, there are constants $a^C_{\max} > a^C_{\min} > 0$ such that

$$a^C_{\min} \le \lambda_{\min}(\mathsf{A}_t) \le \lambda_{\max}(\mathsf{A}_t) \le a^C_{\max} \tag{10.32}$$

for any $t \in \mathscr{T}_C$. As above, we denote by $\lambda_{\min}(\mathsf{A}_t)$ and $\lambda_{\max}(\mathsf{A}_t)$ the extreme eigenvalues of A_t. Our optimality result then reads as follows.

Theorem 10.2 *Let $C \ge 2$, $\rho > 0$, and $\varepsilon > 0$ denote given constants and let $\{\boldsymbol{\lambda}_t^k\}$, $\{\mu_t^k\}$ be generated by Algorithm 9.2 (SMALBE-M) for (10.30) with*

$$\|\mathbf{b}_t\| \ge \eta_t > 0,\ 1 > \beta > 0,\ M_{t,0} > 0, \ \ \mu_t^0 = \mathbf{o}.$$

Let Step 1 of Algorithm 9.2 be implemented by means of Algorithm 8.2 (MPRGP) with parameters $\Gamma > 0$ and $\alpha \in (0, 2/a^C_{\max})$, so that it generates the iterates

$$\boldsymbol{\lambda}_t^{k,0}, \boldsymbol{\lambda}_t^{k,1}, \ldots, \boldsymbol{\lambda}_t^{k,l} = \boldsymbol{\lambda}_t^k$$

for the solution of (10.30) starting from $\boldsymbol{\lambda}_t^{k,0} = \boldsymbol{\lambda}_t^{k-1}$ *with* $\boldsymbol{\lambda}_t^{-1} = \mathbf{o}$*, where* $l = l_{t,k}$ *is the first index satisfying*

$$\|\mathbf{g}^P(\boldsymbol{\lambda}_t^{k,l}, \boldsymbol{\mu}_t^k, \rho_{t,k})\| \le M_{t,k}\|\mathsf{B}_t\boldsymbol{\lambda}_t^{k,l}\| \tag{10.33}$$

or

$$\|\mathbf{g}^P(\boldsymbol{\lambda}_t^{k_t}, \boldsymbol{\mu}_t^{k_t}, \rho_{t,k_t})\| \le \varepsilon\|\mathbf{b}_t\| \quad \textit{and} \quad \|\mathsf{B}_t\boldsymbol{\lambda}_t^{k_t}\| \le \varepsilon\|\mathbf{b}_t\|. \tag{10.34}$$

Then for any $t \in \mathcal{T}_C$ *and problem (10.30), an approximate solution* $\boldsymbol{\lambda}_t^{k_t}$ *which satisfies (10.34) is generated at* $O(1)$ *matrix–vector multiplications by the Hessian* A_t *of* f_t*.*

Proof The class of problems satisfies all assumptions of Theorem 9.4 (i.e., the inequalities (10.31) and (10.32)). The rest follows by Theorem 9.4. □

Since the cost of matrix–vector multiplications by the Hessian A_t is proportional to the number of dual variables, Theorem 10.2 proves the numerical scalability of SMALBE for (10.30) provided the bound constrained minimization in the inner loop is implemented by means of MPRGP. The parallel scalability follows directly from the discussion at the end of Sect. 2.4. See also the next section.

10.9 Numerical Experiments

In this section, we illustrate the numerical scalability of TFETI on the solution of model variational inequalities (10.1).

The domain Ω was first partitioned into identical squares with the side

$$H \in \{1, 1/2, 1/4, 1/8, 1/16, 1/32\}.$$

The square subdomains were then discretized by regular grids with the discretization parameter $h = H/128$, so that the discretized problems have the primal dimension n ranging from 33, 282 to 34,080,768. The computations were performed with the recommended parameters including

$$M_0 = 1, \quad \rho = \|\mathsf{A}_t\|, \quad \Gamma = 1, \quad \text{and} \quad \varepsilon_f = \varepsilon_e = \varepsilon = 10^{-4}.$$

Thus the stopping criterion was

$$\|\mathbf{g}_t^P(\boldsymbol{\lambda}^k)\| \le 10^{-4}\|\mathbf{b}_t\| \quad \text{and} \quad \|\mathsf{B}_t\boldsymbol{\lambda}^k\| \le 10^{-4}\|\mathbf{b}_t\|.$$

The solutions for $H = 1/4$ and $h = 1/4$ are in Figs. 10.4 and 10.5. The results of computations are in Fig. 10.6. We can see that the numbers of matrix–vector multiplications (on vertical axis) vary moderately in agreement with Theorem 10.2 up to a few millions of nodal variables. The reason for the increased number of iterations for large problems is not clear. The latter could have been caused by very fine discretization, unmatched by the discretization of any 3D problem, the increased number of zero active multiplicators, rounding errors, etc.

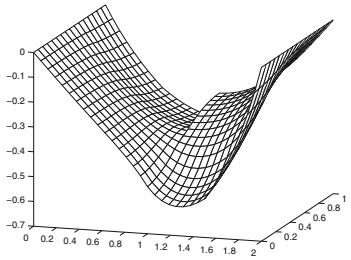

Fig. 10.4 The solution of coercive problem

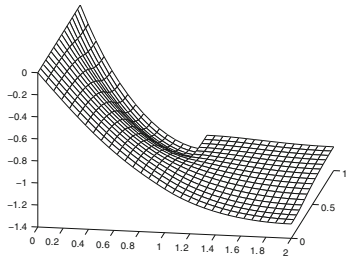

Fig. 10.5 The solution of semicoercive problem

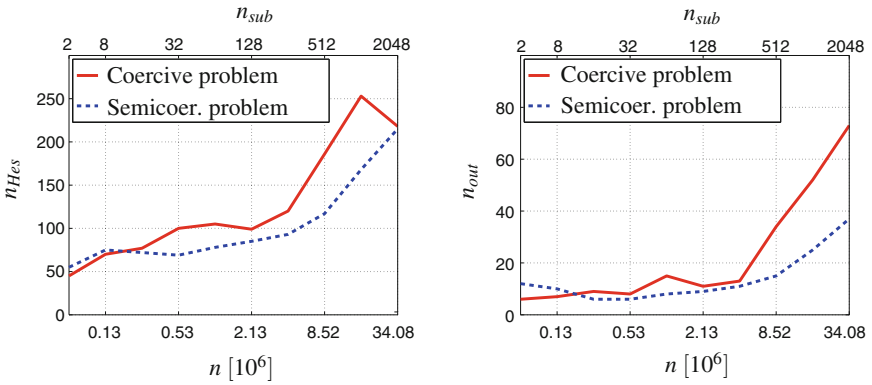

Fig. 10.6 Numerical scalability for the coercive and semicoercive problems. Number of Hessian multiplications n_{Hes} (*left*) and outer iterations n_{out} (*right*)

10.10 Comments and References

The solvability, approximation, and classical numerical methods for elliptic boundary variational inequalities are discussed in the books by Glowinski [10], Glowinski, Lions, and Trèmoliéres [11], and Hlaváček et al. [2]. The dual (reciprocal) formulation of boundary variational inequalities can be found, e.g., in the books by Duvaut and Lions [1] and Hlaváček et al. [2]. More problems described by variational inequalities can be found in Duvaut and Lions [1]. For scalar contact problems, see Sofonea and Matei [12] or Migorski, Ochal, and Sofonea [13].

The first steps toward the development of scalable algorithms for variational inequalities were based on multigrid methods. Hackbusch and Mittelmann [14] and Hoppe [15] used the multigrid to solve auxiliary linear problems and gave a numerical evidence of the efficiency of their algorithms. Using the observations related to Mandel [16], Kornhuber [17] proved the convergence of his monotonic multigrid method. Later overview of multigrid solvers can be found in Gräser and Kornhuber [18].

Our presentation is based on FETI which was proposed by Farhat and Roux [19, 20]. A milestone in the development of domain decomposition algorithms was the proof of numerical scalability of FETI with preconditioning by the "natural coarse grid" by Farhat, Mandel, and Roux [9]. TFETI was proposed for linear problems independently by Dostál, Horák, and Kučera [21] and Of (all floating BETI) [22], [23]. The idea was considered earlier by Park, Felippa, and Gumaste [24].

Augmented Lagrangians for equality constraints were often used to implement active constraints as in Glowinski and LeTallec [25] or Simo and Laursen [26]. The algorithm that combines FETI with the augmented Lagrangians in the outer loop and the inexact solution of bound constrained problems in the inner loop appeared in Dostál, Friedlander, and Santos [27]. See also Dostál, Gomes, and Santos [4, 6]. The first result on the numerical scalability of an algorithm for the solution of a variational inequality used an optimal penalty in dual FETI problem [28]. Using special features of 2D problems, the optimality was proved also for the FETI–DP based algorithm [29, 30] including the solution of coercive problems with non-penetration condition imposed by mortars [31]. An experimental evidence of the scalability of the algorithm with the inner loop implemented by the proportioning [32] was given in Dostál and Horák [33]. The complete proof of optimality that we present here appeared in Dostál and Horák [34].

Other results related to the scalability include Badea, Tai, and Wang [35], who proved the linear rate of convergence for an additive Schwarz domain decomposition method which assumes the exact solution of nonlinear subdomain problems. The variants of two-level FETI methods with preconditioning in face applied to the solution of auxiliary linear problems arising in the solution of the model variational inequality introduced in Sect. 10.1 can be found in Lee [36]. See also the comments in the following chapter.

References

1. Duvaut, G., Lions, J.L.: Inequalities in Mechanics and Physics. Springer, Berlin (1976)
2. Hlaváček, I., Haslinger, J., Nečas, J., Lovíšek, J.: Solution of Variational Inequalities in Mechanics. Springer, Berlin (1988)
3. Glowinski, R.: Numerical Methods for Nonlinear Variational Problems. Springer Series in Computational Physics. Springer, Berlin (1984)
4. Dostál, Z., Gomes, F.A.M., Santos, S.A.: Solution of contact problems by FETI domain decomposition with natural coarse space projection. Comput. Methods Appl. Mech. Eng. **190**(13–14), 1611–1627 (2000)
5. Bochev, P.: On the finite element solution of the pure neumann problem. SIAM Rev. **47**(1), 50–66 (2005)
6. Dostál, Z., Gomes, F.A.M., Santos, S.A.: Duality based domain decomposition with natural coarse space for variational inequalities. J. Comput. Appl. Math. **126**(1–2), 397–415 (2000)
7. Bramble, J.H., Pasciak, J.E., Schatz, A.H.: The construction of preconditioners for elliptic problems by substructuring I. Math. Comput. **47**, 103–134 (1986)
8. Pechstein, C.: Finite and Boundary Element Tearing and Interconnecting Solvers for Multiscale Problems. Sringer, Heidelberg (2013)
9. Farhat, C., Mandel, J., Roux, F.-X.: Optimal convergence properties of the FETI domain decomposition method. Comput. Methods Appl. Mech. Eng. **115**, 365–385 (1994)
10. Glowinski, R.: Variational Inequalities. Springer, Berlin (1980)
11. Glowinski, R., Lions, J.L., Tremolieres, R.: Numerical Analysis of Variational Inequalities. North-Holland, Amsterdam (1981)
12. Sofonea, M., Matei, A.C.: Variational Inequalities with Applications. A Study of Antiplane Frictional Contact Problems. Springer, New York (2009)
13. Migorski, S., Ochal, A., Sofonea, M.: Modeling and analysis of an antiplane piezoelectric contact problem. Math. Models Methods Appl. Sci. **19**, 1295–1324 (2009)
14. Hackbusch, W., Mittelmann, H.: On multi-grid methods for variational inequalities. Numerische Mathematik **42**, 65–76 (1983)
15. Hoppe, R.H.W.: Multigrid algorithms for variational inequalities. SIAM J. Numer. Anal. **24**(5), 1046–1065 (1987)
16. Mandel, J.: Étude algébrique d'une méthode multigrille pour quelques problèmes de frontière libre. Comptes Rendus de l'Académie des Sciences Series **I**(298), 469–472 (1984)
17. Kornhuber, R.: Monotone Multigrid Methods for Nonlinear Variational Problems. Teubner, Stuttgart (1997)
18. Gräser, C., Kornhuber, R.: Multigrid methods for obstacle problems. J. Comput. Math. **27**(1), 1–44 (2009)
19. Farhat, C., Roux, F.-X.: A method of finite element tearing and interconnecting and its parallel solution algorithm. Int. J. Numer. Methods Eng. **32**, 1205–1227 (1991)
20. Farhat, C., Roux, F.-X.: An unconventional domain decomposition method for an efficient parallel solution of large-scale finite element systems. SIAM J. Sci. Comput. **13**, 379–396 (1992)
21. Dostál, Z., Horák, D., Kučera, R.: Total FETI - an easier implementable variant of the FETI method for numerical solution of elliptic PDE. Commun. Numer. Methods Eng. **22**, 1155–1162 (2006)
22. Of, G.: BETI - Gebietszerlegungsmethoden mit schnellen Randelementverfahren und Anwendungen. Ph.D. Thesis, University of Stuttgart (2006)
23. Of, G., Steinbach, O.: The all-floating boundary element tearing and interconnecting method. J. Numer. Math. **17**(4), 277–298 (2009)
24. Park, K.C., Felippa, C.A., Gumaste, U.A.: A localized version of the method of Lagrange multipliers. Comput. Mech. **24**, 476–490 (2000)
25. Glowinski, R., Le Tallec, P.: Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics. SIAM, Philadelphia (1989)

26. Simo, J.C., Laursen, T.A.: An augmented Lagrangian treatment of contact problems involving friction. Comput. Struct. **42**, 97–116 (1992)
27. Dostál, Z., Friedlander, A., Santos, S.A.: Solution of contact problems of elasticity by FETI domain decomposition. Domain decomposition methods 10. AMS, Providence. Contemp. Math. **218**, 82–93 (1998)
28. Dostál, Z., Horák, D.: Scalable FETI with optimal dual penalty for a variational inequality. Numer. Linear Algebra Appl. **11**(5–6), 455–472 (2004)
29. Dostál, Z., Horák, D., Stefanica, D.: A scalable FETI-DP algorithm for a coercive variational inequality. IMACS J. Appl. Numer. Math. **54**(3–4), 378–390 (2005)
30. Dostál, Z., Horák, D., Stefanica, D.: A scalable FETI-DP algorithm for semi-coercive variational inequalities. Comput. Methods Appl. Mech. Eng. **196**(8), 1369–1379 (2007)
31. Dostál, Z., Horák, D., Stefanica, D.: A scalable FETI-DP algorithm with non-penetration mortar conditions on contact interface. J. Comput. Appl. Math. **231**(2), 577–591 (2009)
32. Dostál, Z.: Box constrained quadratic programming with proportioning and projections. SIAM J. Optim. **7**(3), 871–887 (1997)
33. Dostál, Z., Horák, D.: Scalability and FETI based algorithm for large discretized variational inequalities. Math. Comput. Simul. **61**(3–6), 347–357 (2003)
34. Dostál, Z., Horák, D.: Theoretically supported scalable FETI for numerical solution of variational inequalities. SIAM J. Numer. Anal. **45**(2), 500–513 (2007)
35. Badea, L., Tai, X.C., Wang, J.: Convergence rate analysis of a multiplicative Schwarz method for variational inequalities. SIAM J. Numer. Anal. **41**(3), 1052–1073 (2003)
36. Lee, J.: Two domain decomposition methods for auxiliary linear problems for a multibody variational inequality. SIAM J. Sci. Computi. **35**(3), 1350–1375 (2013)

Chapter 11
Frictionless Contact Problems

Now we shall extend the results introduced in the previous chapter to the solution of multibody contact problems of elasticity without friction. We shall restrict our attention to the problems of linear elasticity, i.e., we shall assume small deformations and linear stress-strain relations. Moreover, we shall be interested mainly in computationally challenging 3D problems.

The presentation of TFETI for the solution of frictionless contact problems is very similar to the presentation of TFETI for the solution of scalar variational inequalities in the previous chapter. The main difference, apart from more complicated formulae and kernel spaces, is in the discretization of linearized non-penetration conditions. Here we shall restrict our attention to the most simple node-to-node non-penetration conditions, leaving the discussion of more sophisticated biorthogonal mortars to Chap. 15.

The FETI-type domain decomposition methods comply well with the structure of contact problems, the description of which enhances the decomposition into the subdomains defined by the bodies involved in the problem. Notice that if we decompose the bodies into subdomains, we can view the result as a new multibody problem to find the equilibrium of a system of bodies that are possibly glued and do not penetrate each other. The FETI methods treat each domain separately, which can be effectively exploited in a parallel implementation. Moreover, the algorithm treats very efficiently the "floating" bodies, the Dirichlet boundary conditions of which admit a rigid body motion. A unique feature of FETI is the existence of a projector to the coarse space the complement of which contains the solution. Thus even though the presented methods were developed primarily for the parallel implementation, they are also effective in a sequential implementation.

The basic TFETI-based algorithms presented here can use effectively tens of thousands of cores to solve both coercive and semicoercive contact problems decomposed into tens of thousands of subdomains and discretized by billions of nodal variables. For larger problems, the initialization of the iterative solving procedure, in particular the projectors, starts to dominate the costs. Some modification for emerging exascale technologies are described in Chap. 19.

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3_11

11.1 Linearized Non-penetration Conditions

Let a system of bodies in a reference configuration occupy open bounded domains $\Omega^1, \ldots, \Omega^s \subset \mathbb{R}^3$ with the Lipchitz boundaries $\Gamma^1, \ldots, \Gamma^s$. Suppose that some Γ^p comprises a part $\Gamma_C^{pq} \subseteq \Gamma^p$ that can get into contact with $\overline{\Omega}^q$ as in Fig. 11.1. We assume that Γ_C^p is sufficiently smooth, so that there is a well-defined outer unit normal $\mathbf{n}^p(\mathbf{x})$ at almost each point $\mathbf{x} \in \Gamma_C^p$.

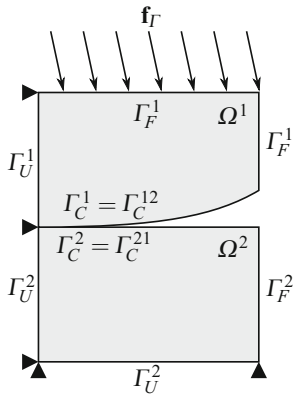

Fig. 11.1 Two-body contact problem

After the deformation, each point $\mathbf{x} \in \Omega^p \cup \Gamma^p$ is transformed into

$$\mathbf{y}^p(\mathbf{x}) = \mathbf{x}^p + \mathbf{u}^p(\mathbf{x}),$$

where $\mathbf{u}^p = \mathbf{u}^p(\mathbf{x}^p)$ is the *displacement vector* which defines the deformation of Ω^p. The mapping $\mathbf{y}^p : \overline{\Omega}^p \to \mathbb{R}^3$ is injective and continuous. The non-penetration condition requires

$$\mathbf{x}^p \in \Gamma_C^p \;\Rightarrow\; \mathbf{x}^p + \mathbf{u}^p(\mathbf{x}^p) \notin \Omega^q, \quad q \in \{1, \ldots, s\}, \quad p \neq q, .$$

It is difficult to enhance the latter condition into an effective computational scheme, so we shall replace it by linearized relations. From each couple $\{p, q\}$ which identify $\Gamma_C^{pq} \neq \emptyset$, we choose one index to identify the *slave side* of a possible contact interface. This choice defines the *contact coupling set* $\mathcal{S}$ of all ordered couples of indices the first component of which refers to the nonempty slave side of the corresponding contact interface. For each $(p, q) \in \mathcal{S}$, we then define a one-to-one continuous mapping

$$\chi^{pq} : \Gamma_C^{pq} \to \Gamma_C^{qp}$$

which assigns to each $\mathbf{x} \in \Gamma_C^{pq}$ a point of the *master side* $\Gamma_C^{qp} \subseteq \Gamma_C^q$ that is assumed to be after the deformation near to $\mathbf{x}$, as in Fig. 11.2. The *(strong) linearized non-penetration condition* then reads

$$\left(\mathbf{u}^p(\mathbf{x}) - \mathbf{u}^q \circ \chi^{pq}(\mathbf{x})\right) \cdot \mathbf{n}^p(\mathbf{x}) \leq \left(\chi^{pq}(\mathbf{x}) - \mathbf{x}\right) \cdot \mathbf{n}^p(\mathbf{x}), \quad \mathbf{x} \in \Gamma_C^{pq}, \quad (p,q) \in \mathscr{S}, \tag{11.1}$$

where $\mathbf{n}^p$ is an approximation of the outer unit normal to Γ^p after the deformation.

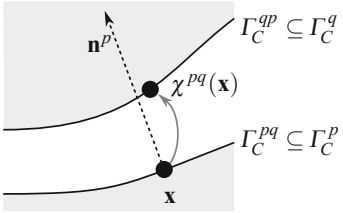

Fig. 11.2 Linearized non-penetration

The linearized condition is exact if $\mathbf{n}^p = \mathbf{n}^p(\mathbf{x})$ is orthogonal to Γ_C^{pq} in the deformed state and the vector $\chi^{pq}(\mathbf{x}) - \mathbf{x}$ moves into the position which is parallel with $\mathbf{n}^p$. This observation can be used to develop an iterative improvement for enforcing the non-penetration condition. See also the discussions in the books by Kikuchi and Oden [1], Laursen [2], or Wriggers [3].

11.2 Equilibrium of a System of Elastic Bodies in Contact

Having described the non-penetration condition, let us switch to the conditions of equilibrium of a system of bodies $\Omega^1, \ldots, \Omega^s$. Let each Γ^p, $p = 1, \ldots, s$, consists of three disjoint parts Γ_U^p, Γ_F^p, and Γ_C^p, $\Gamma^p = \overline{\Gamma}_U^p \cup \overline{\Gamma}_F^p \cup \overline{\Gamma}_C^p$, and let the volume forces $\mathbf{f}^p : \Omega^p \to \mathbb{R}^3$, zero boundary displacements $\mathbf{u}_\Gamma^p : \Gamma_U^p \to \{\mathbf{o}\}$, and the boundary traction $\mathbf{f}_\Gamma^p : \Gamma_F^p \to \mathbb{R}^3$ be given. We admit $\Gamma_U^p = \emptyset$, but in this case we assume some additional restrictions to guarantee that a solution exists. To enhance the contact with a rigid obstacle, we admit the bodies with a priori defined zero displacements. In this case, only the contact boundary of such bodies is relevant in our considerations.

Let us choose a contact coupling set $\mathscr{S}$, so that for each $(p,q) \in \mathscr{S}$, Γ_C^{pq} denotes the part of Γ_C^p which can get into contact with Γ^q, and let us define a one-to-one continuous mapping $\chi^{pq} : \Gamma_C^{pq} \to \Gamma_C^{qp}$ onto the part Γ^{qp} of Γ^q which can come into contact with Γ^p. Thus

$$\overline{\Gamma}_C^p = \cup_{(p,r)\in\mathscr{S}} \overline{\Gamma}_C^{pr} \quad \text{and} \quad \overline{\Gamma}_C^q = \cup_{(r,q)\in\mathscr{S}} \overline{\Gamma}_C^{rq}.$$

Let $\mathbf{v}^p : \Omega^p \cup \Gamma^p \to \mathbb{R}^3,\ p = 1, \ldots, s,$ denote a sufficiently smooth mapping, so that the related concepts are well defined, and denote

$$\mathbf{v} = (\mathbf{v}^1, \ldots, \mathbf{v}^s), \qquad \Omega = \Omega^1 \cup \ldots \Omega^s.$$

Notice that if $\mathbf{x} \in \Omega$, then there is a unique $p = p(\mathbf{x})$ such that $\mathbf{x} \in \Omega^p$, so we can define

$$\mathbf{v}(\mathbf{x}) = \mathbf{v}^{p(\mathbf{x})}(\mathbf{x}) \quad \text{for} \quad \mathbf{x} \in \Omega.$$

We assume that the small strain assumption is satisfied, so that the strain–displacement relations are defined for any $\mathbf{x} \in \Omega$ by *Cauchy's small strain tensor*

$$\varepsilon(\mathbf{v})(\mathbf{x}) = \varepsilon(\mathbf{v}) = 1/2 \left(\nabla \mathbf{v} + (\nabla \mathbf{v})^T\right)$$

with the components

$$e_{ij}(\mathbf{v}) = \frac{1}{2}\left(\frac{\partial v_j}{\partial x_i} + \frac{\partial v_i}{\partial x_j}\right), \quad i, j = 1, 2, 3. \tag{11.2}$$

For simplicity, we assume that the bodies are made of an isotropic linear elastic material so that the constitutive equation for the *Cauchy stress tensor* σ is given in terms of the fourth-order *Hooke elasticity tensor* C by

$$\sigma(\mathbf{v}) = \mathsf{C}\varepsilon(\mathbf{v}) = \lambda \mathrm{tr}(\varepsilon(\mathbf{v}))\mathsf{Id} + 2\mu\varepsilon(\mathbf{v}), \tag{11.3}$$

where $\lambda > 0$ and $\mu > 0$ are the *Lamé parameters* which are assumed to be constant in each subdomain Ω^p, $p = 1, \ldots, s$. The Lamé coefficients can be easily calculated by means of the *Poisson ratio* ν and *Young's modulus* E using

$$\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}, \qquad \mu = \frac{E}{2(1+\nu)},$$

so the components of the elasticity tensor are given by

$$C_{ijk\ell} = \frac{E}{1+\nu}\left(\frac{\nu}{1-2\nu}\delta_{ij}\delta_{k\ell} + \delta_{ik}\delta_{j\ell}\right), \quad i, j, k, \ell = 1, 2, 3. \tag{11.4}$$

The components of the stress tensor are given by

$$\sigma_{ij}(\mathbf{v}) = \sum_{k,\ell=1}^{3} C_{ijk\ell}\, e_{k\ell}(\mathbf{v}), \qquad i, j = 1, 2, 3.$$

Using the above notations, the linearized elastic equilibrium condition and the Dirichlet and Neumann boundary conditions for the displacement $\mathbf{u} = (\mathbf{u}^1, \dots, \mathbf{u}^s)$ can be written as

$$\begin{aligned} -\operatorname{div}\sigma(\mathbf{u}) &= \mathbf{f} \quad \text{in } \Omega, \\ \mathbf{u}^p &= \mathbf{o} \quad \text{on } \Gamma_U^p, \\ \sigma(\mathbf{u}^p)\mathbf{n}^p &= \mathbf{f}_\Gamma^p \text{ on } \Gamma_F^p, \end{aligned} \tag{11.5}$$

where $\mathbf{n}^p$ denotes the outer unit normal to Γ^p which is defined almost everywhere. Here we assume that all objects are sufficiently smooth so that the equations can be satisfied point-wise, postponing more realistic assumptions to the next section. The equations can be written componentwise, e.g., the first equation of (11.5) reads

$$-\sum_{j=1}^{3} \frac{\partial}{\partial x_j}\sigma_{ij}(\mathbf{u}) + f_i = 0 \quad \text{in } \Omega, \ i = 1, 2, 3.$$

To complete the classical formulation of frictionless contact problems, we have to specify the boundary conditions on Γ_C. Assuming that $(p, q) \in \mathscr{S}$, we can use (11.1) and (11.7) to get the non-penetration condition

$$(\mathbf{u} - \mathbf{u} \circ \chi) \cdot \mathbf{n} \le g, \quad \mathbf{x} \in \Gamma_C^p, \quad (p, q) \in \mathscr{S}, \tag{11.6}$$

where we use the notation

$$\begin{aligned} \big(\mathbf{u} - \mathbf{u} \circ \chi\big) \cdot \mathbf{n} &= \big(\mathbf{u}^p(\mathbf{x}) - \mathbf{u}^q \circ \chi^{pq}(\mathbf{x})\big) \cdot \mathbf{n}^p(\mathbf{x}), && \mathbf{x} \in \Gamma_C^{pq}, \\ g &= \big(\chi^{pq}(\mathbf{x}) - \mathbf{x}\big) \cdot \mathbf{n}^p(\mathbf{x}), && \mathbf{x} \in \Gamma_C^{pq}. \end{aligned} \tag{11.7}$$

The surface traction $\boldsymbol{\lambda}$ on the slave side of the active contact interface Γ_C^{pq} and on the rest of Γ_C^{pq} is given by

$$\boldsymbol{\lambda} = \boldsymbol{\lambda}_N = -\sigma(\mathbf{u}^p)\mathbf{n}^p \quad \text{and} \quad \boldsymbol{\lambda} = \mathbf{o},$$

respectively. Since we assume that the contact is frictionless, the tangential component of $\boldsymbol{\lambda}$ is zero, i.e.,

$$\boldsymbol{\lambda} = (\boldsymbol{\lambda} \cdot \mathbf{n}^p)\mathbf{n}^p,$$

and the linearized conditions of equilibrium read

$$\boldsymbol{\lambda} \cdot \mathbf{n}^p \ge 0 \ \text{ and } \ (\boldsymbol{\lambda} \cdot \mathbf{n}^p)\big(\big(\mathbf{u}^p - \mathbf{u}^q \circ \chi\big)\mathbf{n}^p - g\big) = 0, \quad \mathbf{x} \in \Gamma_C^{pq}, \ \ (p, q) \in \mathscr{S}. \tag{11.8}$$

The last condition in (11.8) is called the *complementarity condition*. Newton's law requires that the normal traction acting on the contacting surfaces is equal and opposite, so that

$$-\sigma(\mathbf{u}^q \circ \chi)\mathbf{n}^p = -\boldsymbol{\lambda}, \quad \mathbf{x} \in \Gamma_C^{pq}, \quad (p,q) \in \mathscr{S}. \tag{11.9}$$

The system of equations and inequalities (11.5)–(11.9) with the assumption that the tangential component of $\boldsymbol{\lambda}$ is zero represents the *classical formulation of multi-body frictionless contact problems*. Denoting by λ_n and $[u_n]$ the contact stress and the jump of the boundary displacements, respectively, i.e.,

$$\lambda_n = \boldsymbol{\lambda} \cdot \mathbf{n}^p, \quad [u_n] = \left(\mathbf{u}^p - \mathbf{u}^q \circ \chi\right) \cdot \mathbf{n}^p, \quad \mathbf{x} \in \Gamma_C^{pq},$$

we can write the contact conditions briefly as

$$[u_n] \le g, \ \ \lambda_n \ge 0, \ \ \lambda_n([u_n] - g) = 0, \ \ \boldsymbol{\lambda} = \lambda_n \mathbf{n}^p, \ \ \mathbf{x} \in \Gamma_C^{pq}, \ \ (p,q) \in \mathscr{S}. \tag{11.10}$$

11.3 Variational Formulation

The classical formulation of contact problem (11.5) and (11.10) makes sense only when the solution complies with strong regularity assumptions which are not satisfied by the solution of realistic problems. For example, if a body is not homogeneous, the equilibrium on the material interface requires additional equations.

The basic idea is to require that the equilibrium conditions are satisfied in some average. To formulate it more clearly, let us define the spaces

$$V^p = \left\{\mathbf{v} \in \left(H^1(\Omega^p)\right)^3 : \mathbf{v} = \mathbf{o} \ \text{ on } \ \Gamma_U^p\right\}, \ \ p = 1, \ldots, s, \quad V = V^1 \times \cdots \times V^s,$$

and the convex set

$$\mathscr{K} = \left\{\mathbf{v} \in V : [v_n] \le g \ \text{ on } \ \Gamma_C^{pq}, \ (p,q) \in \mathscr{S}\right\}.$$

Let us first assume that $\mathbf{u}^p, \mathbf{v}^p \in V^p$ are sufficiently smooth, so that we can define

$$\sigma(\mathbf{u}^p) : \varepsilon\left(\mathbf{v}^p\right) = \sum_{i,j=1}^{3} \sigma_{ij}(\mathbf{u}^p)\varepsilon_{ij}\left(\mathbf{v}^p\right)$$

and the bilinear form

$$a^p(\mathbf{u}^p, \mathbf{v}^p) = \int_{\Omega^p} \sigma\left(\mathbf{v}^p\right) : \varepsilon\left(\mathbf{u}^p\right) \, \mathrm{d}\Omega.$$

Using the symmetry of σ, we can get an alternative expression for a^p by means of

$$\begin{aligned} \sigma\left(\mathbf{u}^p\right) : \varepsilon\left(\mathbf{v}^p\right) &= \frac{1}{2} \sum_{i,j=1}^{3} \sigma_{ij}(\mathbf{u}^p) \left(\frac{\partial v_i^p}{\partial x_j} + \frac{\partial v_j^p}{\partial x_i} \right) \\ &= \frac{1}{2} \sum_{i,j=1}^{3} \left(\sigma_{ij}(\mathbf{u}^p) \frac{\partial v_i^p}{\partial x_j} + \sigma_{ji}(\mathbf{v}^p) \frac{\partial v_j^p}{\partial x_i} \right) \\ &= \sum_{i,j=1}^{3} \sigma_{ij}(\mathbf{u}^p) \frac{\partial v_i^p}{\partial x_j} = \sigma(\mathbf{u}^p) : \nabla \mathbf{v}^p. \end{aligned} \tag{11.11}$$

Let us assume that $\mathbf{u} \in \mathscr{K}$ is a sufficiently smooth solution of (11.5) and (11.10), and let $\mathbf{v} \in V$ be sufficiently smooth, so that the Green formula is valid, i.e.,

$$\begin{aligned} \int_{\Omega^p} \sigma(\mathbf{u}^p) : \nabla(\mathbf{v}^p - \mathbf{u}^p) \, \mathrm{d}\Omega = &- \int_{\Omega^p} \operatorname{div} \sigma(\mathbf{u}^p) \cdot (\mathbf{v}^p - \mathbf{u}^p) \, \mathrm{d}\Omega \\ &+ \int_{\Gamma^p} \sigma(\mathbf{u}^p) \mathbf{n}^p \cdot (\mathbf{v}^p - \mathbf{u}^p) \, \mathrm{d}\Gamma. \end{aligned} \tag{11.12}$$

After multiplying the first equation of (11.5) by $\mathbf{v}^p - \mathbf{u}^p$ and integrating the result over Ω^p, we get

$$- \int_{\Omega^p} \operatorname{div} \sigma(\mathbf{u}^p) \cdot (\mathbf{v}^p - \mathbf{u}^p) \, \mathrm{d}\Omega = \int_{\Omega^p} \mathbf{f}^p \cdot (\mathbf{v}^p - \mathbf{u}^p) \, \mathrm{d}\Omega.$$

We can also use (11.11) and (11.12) to get

$$a^p(\mathbf{u}^p, \mathbf{v}^p - \mathbf{u}^p) = - \int_{\Omega^p} \operatorname{div} \sigma(\mathbf{u}^p) \cdot (\mathbf{v}^p - \mathbf{u}^p) \, \mathrm{d}\Omega + \int_{\Gamma^p} \sigma(\mathbf{u}^p) \mathbf{n}^p \cdot (\mathbf{v}^p - \mathbf{u}^p) \, \mathrm{d}\Gamma.$$

After comparing the latter two equations and summing up, we get

$$\sum_{p=1}^{s} a^p(\mathbf{u}^p, \mathbf{v}^p - \mathbf{u}^p) = \sum_{p=1}^{s} \int_{\Omega^p} \mathbf{f}^p \cdot (\mathbf{v}^p - \mathbf{u}^p) \, \mathrm{d}\Omega + \sum_{p=1}^{s} \int_{\Gamma^p} \sigma(\mathbf{u}^p) \mathbf{n}^p \cdot (\mathbf{v}^p - \mathbf{u}^p) \, \mathrm{d}\Gamma.$$

Using the boundary conditions, the boundary integrals can be modified to

$$\int_{\Gamma^p} \sigma(\mathbf{u}^p) \mathbf{n}^p \cdot (\mathbf{v}^p - \mathbf{u}^p) \, \mathrm{d}\Gamma = \int_{\Gamma_F^p} \mathbf{f}_\Gamma^p \cdot (\mathbf{v}^p - \mathbf{u}^p) \, \mathrm{d}\Gamma + \int_{\Gamma_C^p} \sigma(\mathbf{u}^p) \mathbf{n}^p \cdot (\mathbf{v}^p - \mathbf{u}^p) \, \mathrm{d}\Gamma.$$

Denoting

$$a(\mathbf{u},\mathbf{v})=\sum_{p=1}^{s}a^p(\mathbf{u},\mathbf{v}),\quad \ell(\mathbf{v})=\sum_{p=1}^{s}\int_{\Gamma_F^p}\mathbf{f}_\Gamma^p\cdot(\mathbf{v}^p-\mathbf{u}^p)\,\mathrm{d}\Gamma+\sum_{p=1}^{s}\int_{\Omega^p}\mathbf{f}^p\cdot(\mathbf{v}^p-\mathbf{u}^p)\,\mathrm{d}\Omega,$$

we can rewrite the above relations as

$$a(\mathbf{u},\mathbf{v}-\mathbf{u})=\ell(\mathbf{v}-\mathbf{u})+\sum_{p=1}^{s}\int_{\Gamma_C^p}\sigma(\mathbf{u}^p)\mathbf{n}^p.(\mathbf{v}^p-\mathbf{u}^p)\,\mathrm{d}\Gamma. \tag{11.13}$$

Moreover, assuming that $(p,q)\in\mathscr{S}$, $\mathbf{n}^p=-\mathbf{n}^q\circ\chi$, and $\mathbf{v}\in\mathscr{K}$, we get

$$\begin{aligned}\int_{\Gamma_C^{pq}}\sigma(\mathbf{u}^p)\mathbf{n}^p\cdot(\mathbf{v}^p-\mathbf{u}^p)\,\mathrm{d}\Gamma&+\int_{\Gamma_C^{qp}}\sigma(\mathbf{u}^q)\mathbf{n}^q\cdot(\mathbf{v}^q-\mathbf{u}^q)\,\mathrm{d}\Gamma\\&=\int_{\Gamma_C^p}\boldsymbol{\lambda}\cdot\left(\mathbf{u}^p-\mathbf{v}^p+\left(\mathbf{v}^q-\mathbf{u}^q\right)\circ\chi^{pq}\right)\mathrm{d}\Gamma\\&=\int_{\Gamma_C^p}\lambda_n([u_n]-[v_n])\,\mathrm{d}\Gamma\\&=\int_{\Gamma_C^p}\lambda_n([u_n]-g+g-[v_n])\,\mathrm{d}\Gamma\\&=\int_{\Gamma_C^p}\lambda_n(g-[v_n])\,\mathrm{d}\Gamma\geq 0,\end{aligned}$$

so any solution $\mathbf{u}$ of the frictionless problem (11.5) and (11.10) satisfies the variational inequality

$$a(\mathbf{u},\mathbf{v}-\mathbf{u})\geq\ell(\mathbf{v}-\mathbf{u}),\quad \mathbf{v}\in\mathscr{K}. \tag{11.14}$$

Inequality (11.14) characterizes a minimizer of the quadratic function

$$q(\mathbf{v})=\frac{1}{2}a(\mathbf{v},\mathbf{v})-\ell(\mathbf{v})$$

defined on V (Theorem 4.5). Denoting $\mathbf{v}=\mathbf{u}+\mathbf{d}$, we can rewrite (11.14) as

$$a(\mathbf{u},\mathbf{d})-\ell(\mathbf{d})\geq 0,\quad \mathbf{u}+\mathbf{d}\in\mathscr{K},$$

so

$$q(\mathbf{u}+\mathbf{d})-q(\mathbf{u})=a(\mathbf{u},\mathbf{d})-\ell(\mathbf{d})+\frac{1}{2}a(\mathbf{d},\mathbf{d})\geq 0,$$

i.e.,

$$\mathbf{u}=\arg\min_{\mathbf{v}\in\mathscr{K}}q(\mathbf{v}). \tag{11.15}$$

The inequality (11.14) and problem (11.15) are well defined for more general functions than the piece-wise continuously differentiable functions assumed in (11.5) and (11.10). Indeed, if $\mathbf{f}^p \in L^2(\Omega^p)$ and $\mathbf{f}_\Gamma^p \in L^2(\Gamma_F^p)$, then a and ℓ can be evaluated with $\mathbf{v} \in V$ provided the boundary relations concerning $\mathbf{v}^p$ are interpreted in the sense of traces. Moreover, it can be proved that if $\mathbf{u}$ is a solution of such generalized problem, then it has a natural mechanical interpretation.

It is well known that a minimizer of q on $\mathscr{K}$ exists when q is coercive on $\mathscr{K}$, i.e.,

$$\mathbf{v} \in \mathscr{K}, \quad \|\mathbf{v}\| \to \infty \quad \Rightarrow \quad q(\mathbf{v}) \to \infty,$$

with the norm induced by the *broken scalar product*

$$(\mathbf{u}, \mathbf{v}) = \sum_{p=1}^{s} \int_{\Omega^p} \mathbf{u}\mathbf{v} \, \mathrm{d}\Omega .$$

If $\Gamma_U^p = \emptyset$ for some $p \in \{1, \ldots, s\}$, then the coercivity condition is satisfied if

$$a(\mathbf{v}, \mathbf{v}) = 0 \quad \Rightarrow \quad \ell(\mathbf{v}) < 0, \quad \mathbf{v} \in \mathscr{K} . \tag{11.16}$$

In this case a solution exists, but it need not be unique.

11.4 Tearing and Interconnecting

Our next step is to reduce the contact problem into a number of "small" problems at the cost of introducing additional constraints. To this end, let us decompose each Ω^p into subdomains with sufficiently smooth boundaries as in Fig. 11.3, assign each subdomain a unique number, and introduce new "gluing" conditions on the artificial inter-subdomain boundaries. In the early papers, the subdomains were required to be quasi-regular and have similar shape and size. The latter conditions indeed affect the performance of the algorithms, but they are not necessary for the optimality theory.

We decompose appropriately also the parts of the boundaries Γ_U^p, Γ_F^p, and Γ_C^p, $p = 1, \ldots s$, and introduce their numbering to comply with the decomposition of the subdomains. For the artificial inter-subdomain boundaries, we introduce a notation in analogy to that concerning the contact boundary, i.e., Γ_G^{pq} denotes the part of Γ^p which is "glued" to Γ_G^{qp}. Obviously $\Gamma_G^{pq} = \Gamma_G^{qp}$ and it is possible that $\Gamma^{pq} = \emptyset$. We shall also tear the boundary subdomains from the boundary and enforce the Dirichlet boundary conditions by the equality constraints. An auxiliary decomposition of the problem of Fig. 11.1 with renumbered subdomains is in Fig. 11.3.

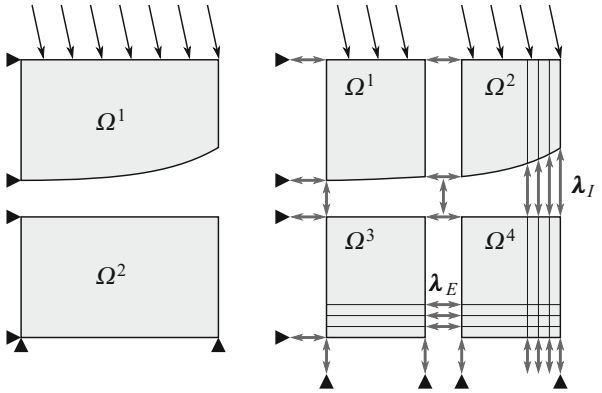

Fig. 11.3 TFETI domain decomposition with subdomain renumbering

To enhance the gluing conditions

$$\mathbf{u}^p = \mathbf{u}^q, \quad \mathbf{x} \in \Gamma_G^{pq}, \tag{11.17}$$

$$\sigma(\mathbf{u}^p)\mathbf{n}^p = -\sigma(\mathbf{u}^q)\mathbf{n}^q, \tag{11.18}$$

into the variational formulation, we shall choose the contact coupling set $\mathscr{S}$ and the test spaces

$$V_{DD}^p = \left\{ \mathbf{v} \in \left(H^1(\Omega^p)\right)^3 : \mathbf{v}^p = \mathbf{o} \text{ on } \Gamma_U^p \right\}, \quad p = 1, \ldots s,$$

$$V_{DD} = V^1 \times \cdots \times V^s,$$

$$\mathscr{K}_{DD} = \left\{ \mathbf{v} \in V_{DD} : [v_n] \le g \text{ on } \Gamma_C^{pq},\ (p,q) \in \mathscr{S};\ \mathbf{v}^p = \mathbf{v}^q \text{ on } \Gamma_G^{pq},\ p,q = 1, \ldots, s. \right\},$$

where the relations should be interpreted in the sense of traces. Recall that for $\mathbf{x} \in \Gamma_C^{pq}$, $(p,q) \in \mathscr{S}$, and $\mathbf{v} \in \mathscr{K}_{DD}$,

$$[v_n] = (\mathbf{v}^p - \mathbf{v}^q \circ \chi^{pq}) \cdot \mathbf{n}^p, \quad g = (\chi^{pq}(\mathbf{x}) - \mathbf{x}) \cdot \mathbf{n}^p.$$

If $\mathbf{u}$ is a classical solution of the decomposed problem which satisfies the gluing conditions (11.17) and (11.18), then for any $\mathbf{v} \in \mathscr{K}_{DD}$ and $p, q = 1, \ldots s$,

$$\int_{\Gamma_G^{pq}} \sigma(\mathbf{u}^p)\mathbf{n}^p \cdot (\mathbf{v}^p - \mathbf{u}^p)\, \mathrm{d}\Gamma + \int_{\Gamma_G^{qp}} \sigma(\mathbf{u}^q)\mathbf{n}^q \cdot (\mathbf{v}^q - \mathbf{u}^q)\, \mathrm{d}\Gamma = 0.$$

It follows that the inequality (11.14) holds also for the forms defined for the decomposed problems. Denoting

$$a(\mathbf{u}, \mathbf{v}) = \sum_{p=1}^{s} a^p(\mathbf{u}^p, \mathbf{v}^p),$$

$$\ell(\mathbf{v}) = \sum_{p=1}^{s} \int_{\Gamma_F^p} \mathbf{f}_\Gamma^p \cdot (\mathbf{v}^p - \mathbf{u}^p)\, \mathrm{d}\Gamma + \sum_{p=1}^{s} \int_{\Omega^p} \mathbf{f}^p \cdot (\mathbf{v}^p - \mathbf{u}^p)\, \mathrm{d}\Omega,$$

we conclude that any classical solution $\mathbf{u}$ of the frictionless decomposed problem (11.5), (11.10), (11.17), and (11.18) satisfies the variational inequality

$$a(\mathbf{u}, \mathbf{v} - \mathbf{u}) \geq \ell(\mathbf{v} - \mathbf{u}), \quad \mathbf{v} \in \mathscr{K}_{DD}, \tag{11.19}$$

and by Theorem 4.5

$$q(\mathbf{u}) \leq q(\mathbf{v}), \quad \mathbf{v} \in \mathscr{K}_{DD}, \quad q(\mathbf{v}) = \frac{1}{2} a(\mathbf{v}, \mathbf{v}) - \ell(\mathbf{v}). \tag{11.20}$$

It can be proved that any sufficiently sooth solution of (11.19) or (11.20) is a classical solution of the frictionless contact problem.

11.5 Discretization

Let us now decompose each subdomain into elements, e.g., tetrahedra, the shape of which is determined by the position of selected nodes (vertices), and let h denote the maximum of the diameters of the elements. We consider such decomposition as a member of the family of elements $\mathscr{T}_h$. We assume that the elements are *shape regular*, i.e., there is a constant $c_s > 0$ independent of h such that the diameter $h(\tau)$ of each element $\tau \in \mathscr{T}^h$ and the radius $\rho(\tau)$ of the largest ball inscribed into τ satisfy

$$\rho(\tau) \geq c_s h(\tau),$$

and that the discretization is *quasi-uniform*, i.e., there is a constant $c_d > 0$ independent of h such that for any element $\tau \in \mathscr{T}^h(\Omega)$

$$h(\tau) \geq c_d\, h.$$

We also assume that the subdomains consist of the unions of elements, i.e., the subdomain boundaries do not cut through any element and the grids are matching on the "gluing" interface of the subdomains. Here we also assume that the grids are matching on the contact interface, i.e., the set of nodes on each slave part Γ_C^{pq} of a

contact interface is mapped by a bijection χ^{pq} onto the set of nodes on the master part Γ_C^{qp} of the contact interface, postponing the generalization to Chap. 15. The finite element approximation of (11.20) gives rise to the QP problem

$$\min \frac{1}{2}\mathbf{u}^T \mathsf{K}\mathbf{u} - \mathbf{f}^T\mathbf{u} \quad \text{subject to} \quad \mathsf{B}_I\mathbf{u} \le \mathbf{c}_I \quad \text{and} \quad \mathsf{B}_E\mathbf{u} = \mathbf{c}_E, \tag{11.21}$$

where

$$\mathsf{K} = \operatorname{diag}(\mathsf{K}_1, \ldots, \mathsf{K}_s)$$

denotes an SPS block-diagonal matrix of order n, B_I denotes an $m_I \times n$ full rank matrix, B_E denotes an $m_E \times n$ full rank matrix, $\mathbf{f} \in \mathbb{R}^n$, $\mathbf{c}_I \in \mathbb{R}^{m_I}$, and $\mathbf{c}_E \in \mathbb{R}^{m_E}$. We use the same notation for nodal displacements as we used for continuous displacements. We shall sometimes denote the nodal displacements by $\mathbf{u}_h$, indicating that it was obtained by the discretization with the finite elements with the diameter less or equal to h.

The blocks K_p, which correspond to Ω^p, are SPS sparse matrices with known kernels, the rigid body modes. Since we consider $3D$ problems, the dimensions of the kernels of K_p and K are six and $6s$, respectively. The vector $\mathbf{f}$ describes the nodal forces arising from the volume forces and/or some other imposed traction.

The matrix $\mathsf{B}_I \in \mathbb{R}^{m_I \times n}$ and the vector $\mathbf{c}_I$ describe the linearized non-penetration conditions. The rows $\mathbf{b}_k$ of B_I are formed by zeros and appropriately placed multiples of coordinates of an approximate outer unit normal on the slave side. If $\mathbf{n}^p$ is an approximate outer normal vector at $\mathbf{x}^p \in \Gamma_C^p$ on the slave side and $\mathbf{x}^q = \chi(\mathbf{x}^p)$ is the corresponding node on the master side, then there is a row $\mathbf{b}_{k*}$ of B_I such that

$$\mathbf{b}_k\mathbf{u}_h = \left(\mathbf{u}_h^p - \mathbf{u}_h^q\right)^T \mathbf{n}^p,$$

where $\mathbf{u}_h^p$ and $\mathbf{u}_h^q$ denote the discretized displacements at $\mathbf{x}^p$ and $\mathbf{x}^q$, respectively. The entry c_k of $\mathbf{c}_I$ describes the normal gap between some $\mathbf{x}^p$ and $\mathbf{x}^q$, i.e.,

$$c_k = (\mathbf{x}^p - \mathbf{x}^q)^T\mathbf{n}^p.$$

The matrix $\mathsf{B}_E \in \mathbb{R}^{m_E \times n}$ enforces the prescribed zero displacements on the part of the boundary with imposed Dirichlet's condition and the continuity of the displacements across the auxiliary interfaces. The continuity requires that $\mathbf{b}_i\mathbf{u}_h = c_i = 0$, where $\mathbf{b}_i$ are the rows of B_E with zero entries except 1 and -1 at appropriate positions. Typically $m = m_I + m_E$ is much smaller than n. If k subdomains have a joint node $\mathbf{x}$, i.e., $\mathbf{x} \in \overline{\Omega}^{i_1} \cap \cdots \cap \overline{\Omega}^{i_k}$, then the gluing of the subdomains at $\mathbf{x}$ for $3D$ problems requires $3(k-1)$ rows of B_E. Notice that the rows of B_E that are associated with different nodes are orthogonal, so that we can use effectively the Gram–Schmidt orthonormalization procedure to get B_E with orthonormal rows.

Remark 11.1 We can achieve that the rows of $\mathsf{B} = [\mathsf{B}_E^T, \mathsf{B}_I^T]^T$ are orthonormal provided each node is involved in at most one inequality. This is always possible for two

bodies or any number of smooth bodies. To simplify the formulation of optimality results, we shall assume in what follows (except Chap. 15) that

$$\mathsf{B}\mathsf{B}^T = \mathsf{I}. \tag{11.22}$$

11.6 Dual Formulation

Even though (11.21) is a standard convex QP problem, its formulation is not suitable for numerical solution. The reasons are that K is typically ill-conditioned, singular, and the feasible set is in general so complex that projections onto it can hardly be effectively computed. Under these circumstances, it would be very difficult to achieve fast identification of the solution active set and to find a fast algorithm for the solution of auxiliary linear problems.

The complications mentioned above can be essentially reduced by applying the duality theory of convex programming (see Sect. 3.7). The Lagrangian associated with problem (11.21) reads

$$L(\mathbf{u}, \boldsymbol{\lambda}_I, \boldsymbol{\lambda}_E) = \frac{1}{2}\mathbf{u}^T\mathsf{K}\mathbf{u} - \mathbf{f}^T\mathbf{u} + \boldsymbol{\lambda}_I^T(\mathsf{B}_I\mathbf{u} - \mathbf{c}_I) + \boldsymbol{\lambda}_E^T(\mathsf{B}_E\mathbf{u} - \mathbf{c}_E), \tag{11.23}$$

where $\boldsymbol{\lambda}_I$ and $\boldsymbol{\lambda}_E$ are the Lagrange multipliers associated with the inequalities and equalities, respectively. Introducing the notation

$$\boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{\lambda}_I \\ \boldsymbol{\lambda}_E \end{bmatrix}, \quad \mathsf{B} = \begin{bmatrix} \mathsf{B}_I \\ \mathsf{B}_E \end{bmatrix}, \quad \text{and} \quad \mathbf{c} = \begin{bmatrix} \mathbf{c}_I \\ \mathbf{c}_E \end{bmatrix},$$

we can write the Lagrangian briefly as

$$L(\mathbf{u}, \boldsymbol{\lambda}) = \frac{1}{2}\mathbf{u}^T\mathsf{K}\mathbf{u} - \mathbf{f}^T\mathbf{u} + \boldsymbol{\lambda}^T(\mathsf{B}\mathbf{u} - \mathbf{c}).$$

Using Proposition 3.13, we get that (11.21) is equivalent to the saddle point problem

$$L(\widehat{\mathbf{u}}, \widehat{\boldsymbol{\lambda}}) = \sup_{\boldsymbol{\lambda}_I \geq \mathbf{o}} \inf_{\mathbf{u}} L(\mathbf{u}, \boldsymbol{\lambda}). \tag{11.24}$$

For a fixed $\boldsymbol{\lambda}$, the Lagrange function $L(\cdot, \boldsymbol{\lambda})$ is convex in the first variable and the minimizer $\mathbf{u}$ of $L(\cdot, \boldsymbol{\lambda})$ satisfies

$$\mathsf{K}\mathbf{u} - \mathbf{f} + \mathsf{B}^T\boldsymbol{\lambda} = \mathbf{o}. \tag{11.25}$$

Equation (11.25) has a solution if and only if

$$\mathbf{f} - \mathsf{B}^T\boldsymbol{\lambda} \in \mathrm{Im}\mathsf{K}, \tag{11.26}$$

which can be expressed more conveniently by means of a matrix R the columns of which span the null space of K as

$$\mathsf{R}^T(\mathbf{f} - \mathsf{B}^T\boldsymbol{\lambda}) = \mathbf{o}. \tag{11.27}$$

The matrix R may be formed directly, block by block, using any basis of the rigid body modes of the subdomains. In our case, each Ω^p is assigned six columns with the blocks

$$\begin{bmatrix} 0 & -z_i & y_i & 1 & 0 & 0 \\ z_i & 0 & -x_i & 0 & 1 & 0 \\ -y_i & x_i & 0 & 0 & 0 & 1 \end{bmatrix}$$

and $\mathsf{O} \in \mathbb{R}^{3\times 6}$ associated with each node $V_i \in \overline{\Omega}^p$ and $V_j \notin \overline{\Omega}^p$, respectively. Using the Gramm–Schmidt procedure, we can find R_i such that

$$\operatorname{Im}\mathsf{R}_i = \operatorname{Ker}\mathsf{K}_i, \quad \mathsf{R} = \operatorname{diag}(\mathsf{R}_1, \ldots, \mathsf{R}_s), \quad \mathsf{R}^T\mathsf{R} = \mathsf{I}.$$

Now assume that $\boldsymbol{\lambda}$ satisfies (11.26), so that we can evaluate $\boldsymbol{\lambda}$ from (11.25) by means of any (left) generalized matrix K^+ which satisfies

$$\mathsf{K}\mathsf{K}^+\mathsf{K} = \mathsf{K}. \tag{11.28}$$

It may be verified directly that if $\mathbf{u}$ solves (11.25), then there is a vector α such that

$$\mathbf{u} = \mathsf{K}^+(\mathbf{f} - \mathsf{B}^T\boldsymbol{\lambda}) + \mathsf{R}\alpha. \tag{11.29}$$

The evaluation of the action of a generalized inverse which satisfies (11.28) is simplified by the block structure of K. Using Lemma 2.1, the kernel of each stiffness matrix K_i can be used to identify a nonsingular submatrix $\mathsf{K}_{\mathscr{J}\mathscr{J}}$ of the same rank as K_i. The action of the left generalized inverse $\mathsf{K}_i^{\#}$ (2.6) can be implemented by Cholesky's decomposition. Observe that

$$\mathsf{K}^{\#} = \operatorname{diag}(\mathsf{K}_1^{\#}, \ldots, \mathsf{K}_s^{\#}).$$

Alternatively, it is possible to use the fixing points strategy of Sect. 11.7.

After substituting expression (11.29) into problem (11.24), changing the signs, and omitting the constant term, we get that $\boldsymbol{\lambda}$ solves the minimization problem

$$\min \Theta(\boldsymbol{\lambda}) \quad \text{s.t.} \quad \boldsymbol{\lambda}_I \geq \mathbf{o} \quad \text{and} \quad \mathsf{R}^T(\mathbf{f} - \mathsf{B}^T\boldsymbol{\lambda}) = \mathbf{o}, \tag{11.30}$$

where

$$\Theta(\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\lambda}^T\mathsf{B}\mathsf{K}^+\mathsf{B}^T\boldsymbol{\lambda} - \boldsymbol{\lambda}^T(\mathsf{B}\mathsf{K}^+\mathbf{f} - \mathbf{c}). \tag{11.31}$$

Once the solution $\widehat{\boldsymbol{\lambda}}$ of (11.30) is known, the solution $\widehat{\mathbf{u}}$ of (11.21) may be evaluated by (11.29) with

$$\alpha = (\mathsf{R}^T\widetilde{\mathsf{B}}^T\widetilde{\mathsf{B}}\mathsf{R})^{-1}\mathsf{R}^T\widetilde{\mathsf{B}}^T(\tilde{\mathbf{c}} - \widetilde{\mathsf{B}}\mathsf{K}^+(\mathbf{f} - \mathsf{B}^T\widehat{\boldsymbol{\lambda}})),$$

where $\widetilde{\mathsf{B}} = \left[\widetilde{\mathsf{B}}_I^T,\ \mathsf{B}_E^T\right]^T$ and $\widetilde{\mathsf{B}}_I$ and $\widetilde{\mathbf{c}_I}$ are formed by the rows of B_I and the components of $\mathbf{c}_I$ that correspond to the positive entries of $\boldsymbol{\lambda}_I$.

11.7 Stable Evaluation of $\mathsf{K}^+\mathbf{x}$ by Using Fixing Nodes

In Sect. 2.4, we described an effective method that can be used for the identification of a nonsingular submatrix $\mathsf{K}_{\mathscr{R}\mathscr{R}}$ of the local stiffness matrix K_i that has its dimension equal to the rank of K_i and used it to the effective evaluation of the action of K^+. Here we shall describe an alternative procedure that can be applied when $\mathsf{K}_{\mathscr{R}\mathscr{R}}$ is ill-conditioned.

To simplify the notation, let us assume that $\mathsf{K} = \mathsf{K}_i \in \mathbb{R}^{n\times n}$. If we choose M nodes that are neither near each other nor placed near any line, $M \geq 3$, then the submatrix $\mathsf{K}_{\mathscr{R}\mathscr{R}}$ of K defined by the set $\mathscr{R}$ of remaining indices is "reasonably" nonsingular. This is not surprising, since $\mathsf{K}_{\mathscr{R}\mathscr{R}}$ is the stiffness matrix of the body that is fixed at the chosen nodes. Using the arguments of mechanics, we deduce that fixing of the chosen nodes makes the body stiff. We call the M chosen nodes the *fixing nodes* and denote by $\mathscr{F}$ the set of indices of the fixed displacements.

We start with the reordering of K to get

$$\widetilde{\mathsf{K}} = \mathsf{P}\mathsf{K}\mathsf{P}^T = \begin{bmatrix} \mathsf{K}_{\mathscr{R}\mathscr{R}} & \mathsf{K}_{\mathscr{R}\mathscr{F}} \\ \mathsf{K}_{\mathscr{F}\mathscr{R}} & \mathsf{K}_{\mathscr{F}\mathscr{F}} \end{bmatrix} = \begin{bmatrix} \mathsf{L}_{\mathscr{R}\mathscr{R}} & \mathsf{O} \\ \mathsf{L}_{\mathscr{F}\mathscr{R}} & \mathsf{I} \end{bmatrix} \begin{bmatrix} \mathsf{L}_{\mathscr{R}\mathscr{R}}^T & \mathsf{L}_{\mathscr{F}\mathscr{R}}^T \\ \mathsf{O} & \mathsf{S} \end{bmatrix}, \tag{11.32}$$

where $\mathsf{L}_{\mathscr{R}\mathscr{R}} \in \mathbb{R}^{r\times r}$ is a lower factor of the Cholesky decomposition of $\mathsf{K}_{\mathscr{R}\mathscr{R}}$, $\mathsf{L}_{\mathscr{F}\mathscr{R}} \in \mathbb{R}^{s\times r}$, $s = 3M$, $\mathsf{L}_{\mathscr{F}\mathscr{R}} = \mathsf{K}_{\mathscr{F}\mathscr{R}}\mathsf{L}_{\mathscr{R}\mathscr{R}}^{-T}$, P is a permutation matrix, and $\mathsf{S} \in \mathbb{R}^{s\times s}$ is the Schur complement matrix of $\widetilde{\mathsf{K}}$ with respect to $\mathsf{K}_{\mathscr{R}\mathscr{R}}$ defined by

$$\mathsf{S} = \mathsf{K}_{\mathscr{F}\mathscr{F}} - \mathsf{K}_{\mathscr{F}\mathscr{R}}\mathsf{K}_{\mathscr{R}\mathscr{R}}^{-1}\mathsf{K}_{\mathscr{R}\mathscr{F}}.$$

To find P, we proceed in two steps. First we form a permutation matrix P_1 to decompose K into blocks

$$\mathsf{P}_1\mathsf{K}\mathsf{P}_1^T = \begin{bmatrix} \mathsf{K}_{\mathscr{R}\mathscr{R}} & \mathsf{K}_{\mathscr{R}\mathscr{F}} \\ \mathsf{K}_{\mathscr{F}\mathscr{R}} & \mathsf{K}_{\mathscr{F}\mathscr{F}} \end{bmatrix}, \tag{11.33}$$

where $\mathsf{K}_{\mathscr{R}\mathscr{R}}$ is nonsingular and $\mathsf{K}_{\mathscr{F}\mathscr{F}}$ corresponds to the degrees of freedom of the M fixing nodes. Then we apply a suitable reordering algorithm on $\mathsf{P}_1\mathsf{K}\mathsf{P}_1^T$ to get a permutation matrix P_2 which leaves the block $\mathsf{K}_{\mathscr{F}\mathscr{F}}$ without changes and enables the sparse Cholesky decomposition of $\mathsf{K}_{\mathscr{R}\mathscr{R}}$. Further, we decompose $\mathsf{P}\mathsf{K}\mathsf{P}^T$ as shown

in (11.32) with $\mathsf{P} = \mathsf{P}_2\mathsf{P}_1$. To preserve the sparsity, we can use any sparse reordering algorithm. The choice depends on the way in which the sparse matrix is stored and on the problem geometry. Using Lemma 2.2, we get

$$\mathsf{K}^+ = \mathsf{P}^T \begin{bmatrix} \mathsf{L}_{\mathscr{R}\mathscr{R}}^{-T} & -\mathsf{L}_{\mathscr{R}\mathscr{R}}^{-T}\mathsf{L}_{\mathscr{F}\mathscr{R}}^{T}\mathsf{S}^+ \\ \mathsf{O} & \mathsf{S}^+ \end{bmatrix} \begin{bmatrix} \mathsf{L}_{\mathscr{R}\mathscr{R}}^{-1} & \mathsf{O} \\ -\mathsf{L}_{\mathscr{F}\mathscr{R}}\mathsf{L}_{\mathscr{R}\mathscr{R}}^{-1} & \mathsf{I} \end{bmatrix} \mathsf{P}, \tag{11.34}$$

where $\mathsf{S}^+ \in \mathbb{R}^{s\times s}$ denotes a left generalized inverse of S.

Since s is small, we can substitute for S^+ the Moore–Penrose generalized inverse $\mathsf{S}^\dagger \in \mathbb{R}^{s\times s}$ computed by SVD. Alternatively, we can use a sparse nonsingular generalized inverse. First observe that the eigenvectors of S that correspond to zero are the traces of vectors from $\operatorname{Ker}\mathsf{K}$ on the fixing nodes. Indeed, if $\widetilde{\mathsf{K}}\mathbf{e} = \mathbf{o}$, then

$$\mathsf{K}_{\mathscr{R}\mathscr{R}}\mathbf{e}_{\mathscr{R}} + \mathsf{K}_{\mathscr{R}\mathscr{F}}\mathbf{e}_{\mathscr{F}} = \mathbf{o}, \qquad \mathsf{K}_{\mathscr{F}\mathscr{R}}\mathbf{e}_{\mathscr{R}} + \mathsf{K}_{\mathscr{F}\mathscr{F}}\mathbf{e}_{\mathscr{F}} = \mathbf{o},$$

and

$$\mathsf{S}\mathbf{e}_{\mathscr{F}} = (\widetilde{\mathsf{K}}_{\mathscr{F}\mathscr{F}} - \widetilde{\mathsf{K}}_{\mathscr{F}\mathscr{R}}\widetilde{\mathsf{K}}_{\mathscr{R}\mathscr{R}}^{-1}\widetilde{\mathsf{K}}_{\mathscr{R}\mathscr{F}})\mathbf{e}_{\mathscr{F}} = \mathbf{o}. \tag{11.35}$$

Having the basis of the kernel od S, we can define the orthogonal projector

$$\mathsf{Q} = \mathsf{R}_{\mathscr{F}*}\left(\mathsf{R}_{\mathscr{F}*}^T\mathsf{R}_{\mathscr{F}*}\right)^{-1}\mathsf{R}_{\mathscr{F}*}^T$$

onto the kernel of S and specify S^+ in (11.34) by

$$\mathsf{S}^+ = (\mathsf{S} + \rho\mathsf{Q})^{-1} = \mathsf{S}^\dagger + \rho^{-1}\mathsf{Q}, \quad \rho > 0.$$

We use $\rho \approx \|\mathsf{K}\|$. To see that S^+ is a left generalized inverse, notice that

$$\mathsf{S}\mathsf{S}^+\mathsf{S} = \mathsf{S}\,(\mathsf{S} + \rho\mathsf{Q})^{-1}\,\mathsf{S} = \mathsf{S}\left(\mathsf{S}^\dagger + \rho^{-1}\mathsf{Q}\right)\mathsf{S} = \mathsf{S}\mathsf{S}^\dagger\mathsf{S} + \rho^{-1}\mathsf{S}\mathsf{Q}\mathsf{S} = \mathsf{S}.$$

Such approach can be considered as a variant of regularization [4].

To implement the above-mentioned observations, it is necessary to have an effective procedure for choosing uniformly distributed fixing nodes. Here we describe a simple but effective method that combines a mesh partitioning algorithm with a method for finding a mesh center. The algorithm reads as follows.

Algorithm 11.1 Algorithm for finding M uniformly distributed fixing nodes in the graph of the discretization.

Given a mesh and $M > 0$.
Step 1. Split the mesh into M submeshes using the mesh partitioning algorithm.
Step 2. Verify whether the resulting submeshes are connected. If not, use a graph postprocessing to get connected submeshes.
Step 3. In each submesh, choose a node which is near its center.

Step 1 can be carried out by any code for graph decompositions such as METIS [5]. The implementation of Step 3 can be based on Corollary 2.1. If the mesh is approximately regular, we can expect that more walks of length k originate from the nodes that are near a center of the mesh. It simply follows that the node with the index i which satisfies

$$w(i,k) \geq w(j,k),\ j = 1,2,\ldots,n,$$

for sufficiently large k is in a sense near to the center of the mesh and can be used to implement Step 3 of Algorithm 11.1.

Recall that the vector

$$\mathbf{p} = \lim_{k\to\infty} \|\mathsf{D}^k\mathbf{e}\|^{-1}\mathsf{D}^k\mathbf{e}$$

is a unique nonnegative eigenvector which corresponds to the largest eigenvalue of the mesh adjacency matrix D. It is also known as the Perron vector of D [6]. It can be approximated by a few steps of the Lanczos method [7]. Thus the index of an approximation of the component of the Perron vector of D is a good approximation of the center of the graph of triangulation. See Brzobohatý et al. [8] for more details and illustrations of the effect of the choice of fixing nodes on the conditioning of the generalized inverses.

11.8 Preconditioning by Projectors to Rigid Body Modes

As in Chap. 10, further improvement can be achieved by adapting some simple observations originating in Farhat, Mandel, and Roux [9]. Let us first simplify the notations by denoting

$$\begin{aligned} \widetilde{\mathsf{F}} &= \mathsf{B}\mathsf{K}^{+}\mathsf{B}^T, & F &= \|\widetilde{\mathsf{F}}\| \\ \mathsf{F} &= F^{-1}\widetilde{\mathsf{F}}, & \widetilde{\mathbf{d}} &= F^{-1}(\mathsf{B}\mathsf{K}^{+}\mathbf{f} - \mathbf{c}), \\ \widetilde{\mathsf{G}} &= \mathsf{R}^T\mathsf{B}^T, & \widetilde{\mathbf{e}} &= \mathsf{R}^T\mathbf{f}, \end{aligned}$$

and let T denote a regular matrix which defines the orthonormalization of the rows of $\widetilde{\mathsf{G}}$ so that the matrix

$$\mathsf{G} = \mathsf{T}\widetilde{\mathsf{G}}$$

has orthonormal rows. After denoting

$$\mathbf{e} = \mathsf{T}\widetilde{\mathbf{e}},$$

problem (11.30) reads

$$\min \quad \frac{1}{2}\boldsymbol{\lambda}^T\mathsf{F}\boldsymbol{\lambda} - \boldsymbol{\lambda}^T\widetilde{\mathbf{d}} \quad \text{s.t.} \quad \boldsymbol{\lambda}_I \geq \mathbf{o} \quad \text{and} \quad \mathsf{G}\boldsymbol{\lambda} = \mathbf{e}. \tag{11.36}$$

For practical computation, we can use an approximate value of the norm

$$F \approx \|\widetilde{\mathsf{F}}\|.$$

Notice that a good approximation of $\|\widetilde{\mathsf{F}}\|$ can be obtained by substituting the estimates of $\|\mathsf{S}_i^{-1}\|$ into (10.27). The scaling of F was not necessary in Sect. 10.6.

As in Chap. 10, we shall transform the problem of minimization on the subset of an affine space to that on the subset of a vector space by means of arbitrary $\widetilde{\boldsymbol{\lambda}}$ which satisfies

$$\mathsf{G}\widetilde{\boldsymbol{\lambda}} = \mathbf{e}. \tag{11.37}$$

Having $\widetilde{\boldsymbol{\lambda}}$, we can look for the solution of (11.36) in the form $\boldsymbol{\lambda} = \boldsymbol{\mu} + \widetilde{\boldsymbol{\lambda}}$. Though a natural choice for $\widetilde{\boldsymbol{\lambda}}$ is the least squares solution

$$\widetilde{\boldsymbol{\lambda}} = \mathsf{G}^T(\mathsf{G}\mathsf{G}^T)^{-1}\mathbf{e},$$

the following lemma shows that if we solve a coercive problem (11.20), then we can even find $\widetilde{\boldsymbol{\lambda}}$ such that $\widetilde{\boldsymbol{\lambda}}_I = \mathbf{o}$.

Lemma 11.1 *Let the problem (11.21) be obtained by the discretization of a coercive problem, i.e., let the prescribed displacements of each body Ω^p be sufficient to prevent its rigid body motion and let $\mathsf{G} = [\mathsf{G}_I, \mathsf{G}_E]$. Then G_E is a full rank matrix and*

$$\widetilde{\boldsymbol{\lambda}} = \begin{bmatrix} \mathbf{o}_I \\ \mathsf{G}_E^T(\mathsf{G}_E\mathsf{G}_E^T)^{-1}\mathbf{e} \end{bmatrix} \tag{11.38}$$

satisfies $\widetilde{\boldsymbol{\lambda}}_I = \mathbf{o}_I$ and $\mathsf{G}\widetilde{\boldsymbol{\lambda}} = \mathbf{e}$.

Proof First observe that

$$\mathsf{G} = [\mathsf{T}\widetilde{\mathsf{G}}_I, \mathsf{T}\widetilde{\mathsf{G}}_E],$$

so it is enough to prove that $\widetilde{\mathsf{G}}_E^T\boldsymbol{\xi} = \mathsf{B}_E\mathsf{R}\boldsymbol{\xi} = \mathbf{o}$ implies $\boldsymbol{\xi} = \mathbf{o}$. Since the entries of $\mathsf{B}_E\mathsf{R}\boldsymbol{\xi}$ denote the jumps of $\mathsf{R}\boldsymbol{\xi}$ across the auxiliary interfaces or the violation of prescribed Dirichlet boundary conditions, it follows that $\mathsf{B}_E\mathsf{R}\boldsymbol{\xi} = \mathbf{o}$ implies that $\mathbf{u} = \mathsf{R}\boldsymbol{\xi}$ satisfies both the discretized Dirichlet conditions and the "gluing" conditions, but belongs to the kernel of K. Thus $\boldsymbol{\xi} \neq \mathbf{o}$ contradicts our assumption that the problem (11.21) was obtained by a correct discretization of (11.20). □

Let us point out that the choice of $\widetilde{\boldsymbol{\lambda}}$ based on Lemma 11.1 guarantees $\mathbf{o}$ to be a feasible vector of the homogenized problem. If the problem (11.20) is semicoercive, we can achieve the same effect with $\widetilde{\boldsymbol{\lambda}}$ which solves

$$\min \frac{1}{2}\|\boldsymbol{\lambda}\|^2 \quad \text{subject to} \quad \mathsf{G}\boldsymbol{\lambda} = \mathbf{e} \ \text{ and } \ \boldsymbol{\lambda} \geq \mathbf{o}.$$

To carry out the transformation, denote $\boldsymbol{\lambda} = \boldsymbol{\mu} + \widetilde{\boldsymbol{\lambda}}$, so that

$$\frac{1}{2}\boldsymbol{\lambda}^T \mathsf{F}\boldsymbol{\lambda} - \boldsymbol{\lambda}^T \widetilde{\mathbf{d}} = \frac{1}{2}\boldsymbol{\mu}^T \mathsf{F}\boldsymbol{\mu} - \boldsymbol{\mu}^T(\widetilde{\mathbf{d}} - \mathsf{F}\widetilde{\boldsymbol{\lambda}}) + \frac{1}{2}\widetilde{\boldsymbol{\lambda}}^T \mathsf{F}\widetilde{\boldsymbol{\lambda}} - \widetilde{\boldsymbol{\lambda}}^T \widetilde{\mathbf{d}}$$

and problem (11.36) has, after returning to the old notation, the same solution as

$$\min \theta(\boldsymbol{\lambda}) \quad \text{s.t.} \quad \mathsf{G}\boldsymbol{\lambda} = \mathbf{o} \quad \text{and} \quad \boldsymbol{\lambda}_I \geq \boldsymbol{\ell}_I = -\widetilde{\boldsymbol{\lambda}}_I \tag{11.39}$$

with

$$\theta(\lambda) = \frac{1}{2}\boldsymbol{\lambda}^T \mathsf{F}\boldsymbol{\lambda} - \boldsymbol{\lambda}^T \mathbf{d}, \quad \mathbf{d} = \widetilde{\mathbf{d}} - \mathsf{F}\widetilde{\boldsymbol{\lambda}}.$$

Let us point out that the proof of Lemma 11.1 is the only place where we need the assumption that our problem (11.21) is coercive.

Our final step is based on the observation that problem (11.39) is equivalent to

$$\min \overline{\theta}_\rho(\lambda) \quad \text{s.t.} \quad \mathsf{G}\boldsymbol{\lambda} = \mathbf{o} \quad \text{and} \quad \boldsymbol{\lambda}_I \geq \boldsymbol{\ell}_I, \tag{11.40}$$

where ρ is an arbitrary positive constant,

$$\overline{\theta}_\rho(\lambda) = \frac{1}{2}\boldsymbol{\lambda}^T(\mathsf{PFP} + \rho\mathsf{Q})\boldsymbol{\lambda} - \boldsymbol{\lambda}^T \mathsf{P}\mathbf{d}, \quad \mathsf{Q} = \mathsf{G}^T\mathsf{G}, \quad \text{and} \quad \mathsf{P} = \mathsf{I} - \mathsf{Q}.$$

Recall that P and Q denote the orthogonal projectors on the image space of G^T and on the kernel of G, respectively. The regularization term is introduced in order to simplify the reference to the results of QP that assume the nonsingular Hessian of f.

11.9 Bounds on the Spectrum

Problem (11.40) turns out to be a suitable starting point for the development of an efficient algorithm for variational inequalities due to the favorable distribution of the spectrum of the Hessian $\mathsf{H} = \mathsf{PFP} + \rho\mathsf{Q}$ of the cost function $\overline{\theta}$. The following lemma, the variant of Lemma 10.3 for linear elasticity, is important for the analysis of optimality of the presented algorithms.

Lemma 11.2 *Let H denote the diameter of a homogeneous body which occupies a domain $\Omega \subset \mathbb{R}^3$, let $\mathsf{K}_{H,h}$ denote its stiffness matrix obtained by the quasi-uniform discretization with shape regular elements of the diameter less or equal to h, and let $\mathsf{S}_{H,h}$ denote the Schur complement of $\mathsf{K}_{H,h}$ with respect to the interior nodes. Let the constraint matrix B satisfy (11.22).*

Then there are constants c and C independent of h and H such that

$$c\frac{h^2}{H}\|\boldsymbol{\lambda}\|^2 \leq \boldsymbol{\lambda}^T \mathsf{S}_{H,h}\boldsymbol{\lambda} \leq Ch\|\boldsymbol{\lambda}\|^2, \quad \boldsymbol{\lambda} \in \mathrm{Im}\mathsf{S}. \tag{11.41}$$

Proof The proof of Lemma 11.2 is similar to the proof of Lemma 10.3. The proof of the upper bound follows from Lemma 13.1 and the obvious inequality

$$\|\mathsf{S}_{H,h}\| \leq \|\mathsf{K}_{H,h}\|. \qquad \square$$

The following theorem is now an easy corollary of Lemmas 10.2 and 11.2.

Theorem 11.1 *Let $\rho > 0$ and let $\mathsf{H}_{\rho,H,h}$ denote the Hessian of $\overline{\theta}_\rho$ resulting from the decomposition and quasi-uniform finite element discretization of problem (11.20) with shape regular elements using the decomposition and discretization parameters H and h, respectively. Let (11.22) be true.*

Then there are constants c and C independent of h and H such that

$$c\|\boldsymbol{\lambda}\|^2 \leq \boldsymbol{\lambda}^T \mathsf{H}_{\rho,H,h}\boldsymbol{\lambda} \leq C\frac{H}{h}\|\boldsymbol{\lambda}\|^2, \quad \boldsymbol{\lambda} \in \mathbb{R}^m. \tag{11.42}$$

Proof Substitute (11.41) into Lemma 10.2 and take into account the regularization term $\rho\mathsf{Q}$ with fixed ρ and the scaling of F.

Remark 11.2 A variant of the theorem was a key ingredient of the first optimality analysis of FETI for linear problems by Farhat, Mandel, and Roux [9].

11.10 Optimality

To show that Algorithm 9.2 (SMALBE-M) with the inner loop implemented by Algorithm 8.2 (MPRGP) is optimal for the solution of a class of problems arising from varying discretizations of a given frictionless contact problem, let us introduce a new notation that complies with that used in the analysis of the algorithms in Part II.

Let $\rho > 0$ and $C \geq 2$ denote given constants and let

$$\mathscr{T}_C = \{(H,h) \in \mathbb{R}^2 : \ H/h \leq C\}$$

denote the set of indices. For any $t \in \mathscr{T}_C$, let us define

$$\begin{aligned} \mathsf{A}_t &= \mathsf{PFP} + \rho\mathsf{Q}, \quad & \mathbf{b}_t &= \mathsf{P}\mathbf{d}, \\ \mathsf{B}_t &= \mathsf{G}, & \boldsymbol{\ell}_I^t &= -\widetilde{\boldsymbol{\lambda}}_I, \end{aligned}$$

where the vectors and matrices are those arising from the discretization of (11.20) with $t = (H,h)$. We assume that the discretization satisfies the assumptions of Theorem 11.1 and $\boldsymbol{\ell}_I^t \leq \mathbf{o}$. We get a class of problems

$$\min f_t(\boldsymbol{\lambda}) \ \text{ subject to } \ \mathsf{B}_t\boldsymbol{\lambda} = \mathbf{o} \ \text{ and } \ \boldsymbol{\lambda}_I \geq \ell_I^t \tag{11.43}$$

with

$$f_t(\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\lambda}^T \mathsf{A}_t \boldsymbol{\lambda} - \mathbf{b}_t^T \boldsymbol{\lambda}.$$

Using these definitions and $\mathsf{G}\mathsf{G}^T = \mathsf{I}$, we obtain

$$\|\mathsf{B}_t\| \le 1. \tag{11.44}$$

It follows by Theorem 11.1 that for any $C \ge 2$ there are constants

$$a^C_{\max} > a^C_{\min} > 0$$

such that

$$a^C_{\min} \le \alpha_{\min}(\mathsf{A}_t) \le \alpha_{\max}(\mathsf{A}_t) \le a^C_{\max} \tag{11.45}$$

for any $t \in \mathcal{T}_C$. In particular, it follows that the assumptions of Theorem 9.4 are satisfied for any set of indices $\mathcal{T}_C$, $C \ge 2$, and we can formulate the main result of this chapter.

Theorem 11.2 *Let $C \ge 2$, $\rho > 0$, and $\varepsilon > 0$ denote given constants and let $\{\boldsymbol{\lambda}_t^k\}$, $\{\boldsymbol{\mu}_t^k\}$, and $\{M_{t,k}\}$ be generated by Algorithm 9.2 (SMALBE-M) for (11.43) with*

$$\|\mathbf{b}_t\| \ge \eta_t > 0, \quad 1 > \beta > 0, \quad M_{t,0} = M_0 > 0, \quad \rho > 0, \quad \boldsymbol{\mu}_t^0 = \mathbf{o}.$$

Let Step 1 of Algorithm 9.2 be implemented by means of Algorithm 8.2 (MPRGP) with parameters

$$\Gamma > 0 \text{ and } \alpha \in (0, 2/a^C_{\max}),$$

so that it generates the iterates

$$\boldsymbol{\lambda}_t^{k,0}, \boldsymbol{\lambda}_t^{k,1}, \ldots, \boldsymbol{\lambda}_t^{k,l} = \boldsymbol{\lambda}_t^k$$

for the solution of (11.43) starting from $\boldsymbol{\lambda}_t^{k,0} = \boldsymbol{\lambda}_t^{k-1}$ with $\boldsymbol{\lambda}_t^{-1} = \mathbf{o}$, where $l = l_{t,k}$ is the first index satisfying

$$\|\mathbf{g}^P(\boldsymbol{\lambda}_t^{k,l}, \boldsymbol{\mu}_t^k, \rho)\| \le M_{t,k}\|\mathsf{B}_t \boldsymbol{\lambda}_t^{k,l}\| \tag{11.46}$$

or

$$\|\mathbf{g}^P(\boldsymbol{\lambda}_t^{k,l}, \boldsymbol{\mu}_t^k, \rho)\| \le \varepsilon\|\mathbf{b}_t\| \quad \text{and} \quad \|\mathsf{B}_t \boldsymbol{\lambda}_t^{k,l}\| \le \varepsilon\|\mathbf{b}_t\|. \tag{11.47}$$

Then for any $t \in \mathcal{T}_C$ and problem (11.43), Algorithm 9.2 generates an approximate solution $\boldsymbol{\lambda}_t^{k_t}$ which satisfies (11.47) at $O(1)$ matrix–vector multiplications by the Hessian A_t of f_t.

11.11 Numerical Experiments

The algorithms presented here were implemented in several software packages (see Sect. 19.5) and tested on a number of academic benchmarks and real-world problems. Here we give some results that illustrate their numerical scalability and effectiveness using `MatSol` (see Sect. 19.5.1 or Kozubek et al. [10]), postponing the demonstration of parallel scalability to Chap. 19. All the computations were carried out with the parameters recommended in the description of the algorithms in Chaps. 7–9. The relative precision of the computations was $\varepsilon = 10^{-4}$ (see (9.40)).

11.11.1 Academic Benchmark

Let us consider a 3D semicoercive contact problem of two cantilever beams of sizes $2 \times 1 \times 1$ [m] in mutual contact without friction. The beams are depicted in Fig. 11.4. Zero horizontal displacements were prescribed on right face of the upper beam. The lower beam (shifted by 1 [m]) was fixed on its left face. The vertical traction $\mathbf{f} = 20$ [MPa] was prescribed on the upper and left faces of the upper beam.

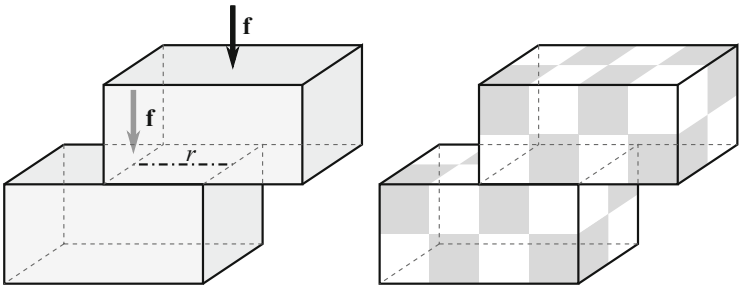

Fig. 11.4 Two beams benchmark and its decomposition

The problem was discretized with varying discretization and decomposition parameters h and H, respectively. For each h and H, the bodies were decomposed into $2/H \times 1/H \times 1/H$ subdomains discretized by hexahedral elements. We kept $H/h = 8$, so that the assumptions of Theorem 11.2 were satisfied.

The performance of the algorithms is documented in the following graphs. The numbers n_{out} of the outer iterations of SMALBE and n_{Hes} of the multiplication by the Hessian F of the dual function depending on the primal dimension n is depicted in Fig. 11.5. We can see stable numbers of both inner and outer iterations for n ranging from 431,244 to 11,643,588. The dual dimension of the problems ranged from 88,601 to 2,728,955. We conclude that the performance of the algorithm is in agreement with the theory.

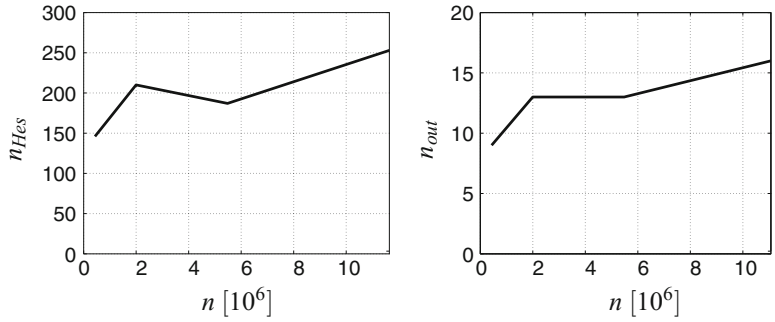

Fig. 11.5 Cantilever beams—numbers of matrix—vector multiplications by F (*left*) and outer iterations (*right*)

The normal traction along the axis of the contact interface is in Fig. 11.6. Figure 11.6 shows that in the solution of the largest problem, most of 11,550 linear inequality constraints were active.

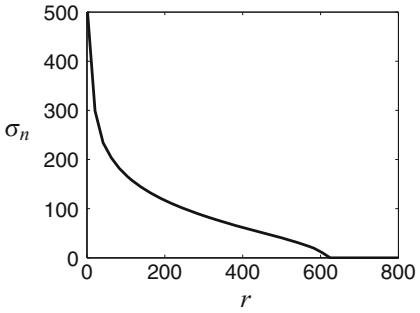

Fig. 11.6 Normal contact pressures along the line r

11.11.2 Roller Bearings of Wind Generator

We have also tested our algorithms on real-world problems, including the stress analysis in the roller bearings of a wind generator that is depicted in Fig. 11.7. The problem is difficult because it comprises 73 bodies in mutual contact and only one is fixed in space.

Fig. 11.7 Frictionless roller bearing of wind generator

The solution of the problem discretized by 2,730,000/459,800 primal/dual variables and decomposed into 700 subdomains required 4270 matrix–vector multiplications, including outer iterations for exact non-penetration. The von Mises stress distribution is in Fig. 11.7 (right). Though the number of iterations is not small, the parallel scalability of the algorithm enables to obtain the solution in a reasonable time.

11.12 Comments and References

A convenient presentation of the variational formulation, including the dual formulation, the results concerning the existence and uniqueness of a solution, the finite element approximation of the solution, and standard iterative methods for the solution can be found in the book by Kikuchi and Oden [1]. For the variational formulation and analysis, see also Hlaváček et al. [11]. The up-to-date engineering approach to the solution of contact problems can be found in Laursen [2] or Wriggers [3]. See Chap. 15 for the discussion of combination of TFETI and mortar approximation of contact conditions.

Probably the first theoretical results concerning the development of scalable algorithms for coercive contact problems were proved by Schöberl [12, 13]. A numerical evidence of scalability of a different approach combining FETI–DP with a Newton-type algorithm and preconditioning in face by standard FETI preconditioners for 3D contact problems was given in Duresseix and Farhat [14] and Avery et al. [15]. See also Dostál et al. [16]. Impressive applications of the above approach can be found in [17].

A stable implementation of TFETI requires reliable evaluation of the action of a generalized inverse of the SPS stiffness matrix. The presentation in Sect. 11.6 combines the earlier observations by Savenkov, Andrä, and Iliev [4] and Felippa and Park [18] on the regularization and Farhat and Géradin [19] on the application of the LU and SVD decompositions. Our exposition uses the fixing nodes presented in Brzobohatý et al. [8].

It should be noted that the effort to develop scalable solvers for coercive variational inequalities was not restricted to FETI. Optimal properties of multigrid methods for linear problems were exploited, e.g., by Kornhuber and Krause [20] and Kornhuber et al. [21], to give an experimental evidence of the numerical scalability of an algorithm based on monotonic multigrid. However, as pointed out by Iontcheva and Vassilevski [22], the coarse grid used in the multigrid should avoid the constrained variables, so that its approximation properties are limited and not sufficient to support the proof of optimality of the nonlinear algorithm. Multigrid has been used also in the framework of the nonsmooth Newton methods, which turned out to be an especially effective tool for the solution of problems with complex nonlinearities (see Sect. 12.10).

The augmented Lagrangians were often used in engineering algorithms to implement equality constraints as in Simo and Laursen [23] or Glowinski and Le Tallec [24]. It seems that the first application of the LANCELOT style [25] augmented Lagrangians (proposed for bound and general equality constrains) with adaptive precision control in combination with FETI to the solution of contact problems is in Dostál, Friedlander, and Santos [26] and Dostál, Gomes, and Santos [27, 28]. The experimental evidence of numerical scalability was presented in Dostál et al. [29]. The optimality was proved in [30]—the proof exploits the optimal properties of MPRGP [31] (see also Sect. 9.10), SMALBE-M (see [32, 33], or Sect. 8), and TFETI (see [34]). Here we partly follow [30].

The linear steps of MPRGP can be preconditioned by the standard FETI preconditioners, i.e., the lumped preconditioner or Dirichlet's preconditioner [35]. The preconditioning by the conjugate projector for the FETI–DP solution of contact problem was presented by Jarošová, Klawonn, and Rheinbach [36]. However, our experience does not indicate high efficiency of the modified algorithms for contact problems. The negative effect of jumping coefficients can be reduced by the reorthogonalization based preconditioning or renormalization based scaling presented in Chap. 16.

There is an interesting corollary of our theory. If we are given a class of contact problems which involves the bodies that are discretized by quasi-uniform grids using shape regular elements, so that the regular part of their spectrum is contained in a given positive interval, then Theorem 11.2 implies that in spite of nonlinearity, *there is a bound, independent of a number of the bodies, on the number of iterations that are necessary to approximate the solution to a given precision.*

References

1. Kikuchi, N., Oden, J.T.: Contact Problems in Elasticity. SIAM, Philadelphia (1988)
2. Laursen, T.: Computational Contact and Impact Mechanics. Springer, Berlin (2002)
3. Wriggers, P.: Contact Mechanics. Springer, Berlin (2005)
4. Savenkov, E., Andrä, H., Iliev, O.: An analysis of one regularization approach for solution of pure Neumann problem. Berichte des Faruenhofer ITWM, Nr. 137, Kaiserslautern (2008)

5. Karypis, G.: METIS – a family of programs for partitioning unstructured graphs and hypergraphs and computing fill-reducing orderings of sparse matrices. http://glaros.dtc.umn.edu/gkhome/views/metis
6. Berman, A., Plemmons, R.J.: Nonnegative Matrices in the Mathematical Sciences. SIAM, Philadelphia (1994)
7. Golub, G.H., Van Loan, C.F.: Matrix Computations, 2nd edn. Johns Hopkins University Press, Baltimore (1989)
8. Brzobohatý, T., Dostál, Z., Kozubek, T., Kovář, P., Markopoulos, A.: Cholesky decomposition with fixing nodes to stable computation of a generalized inverse of the stiffness matrix of a floating structure. Int. J. Numer. Methods Eng. **88**(5), 493–509 (2011)
9. Farhat, C., Mandel, J., Roux, F.-X.: Optimal convergence properties of the FETI domain decomposition method. Comput. Methods Appl. Mech. Eng. **115**, 365–385 (1994)
10. Kozubek, T., Markopoulos, A., Brzobohatý, T., Kučera, R., Vondrák, V., Dostál, Z.: MatSol–MATLAB efficient solvers for problems in engineering (2015). http://industry.it4i.cz/en/products/matsol/
11. Hlaváček, I., Haslinger, J., Nečas, J., Lovíšek, J.: Solution of Variational Inequalities in Mechanics. Springer, Berlin (1988)
12. Schöberl, J.: Solving the Signorini problem on the basis of domain decomposition techniques. Computing **60**(4), 323–344 (1998)
13. Schöberl, J.: Efficient contact solvers based on domain decomposition techniques. Comput. Math. Appl. **42**, 1217–1228 (2001)
14. Dureisseix, D., Farhat, C.: A numerically scalable domain decomposition method for solution of frictionless contact problems. Int. J. Numer. Methods Eng. **50**(12), 2643–2666 (2001)
15. Avery, P., Rebel, G., Lesoinne, M., Farhat, C.: A numerically scalable dual-primal substructuring method for the solution of contact problems - part I: the frictionless case. Comput. Methods Appl. Mech. Eng. **193**, 2403–2426 (2004)
16. Dostál, Z., Vondrák, V., Horák, D., Farhat, C., Avery, P.: Scalable FETI algorithms for frictionless contact problems. Lecture Notes in Computational Science and Engineering, vol. 60, pp. 263–270. Springer, Berlin (2008)
17. Avery, P., Farhat, C.: The FETI family of domain decomposition methods for inequality-constrained quadratic programming: application to contact problems with conforming and nonconforming interfaces. Comput. Methods Appl. Mech. Eng. **198**, 1673–1683 (2009)
18. Felippa, C.A., Park, K.C.: The construction of free-free flexibility matrices for multilevel structural analysis. Comput. Methods Appl. Mech. Eng. **191**, 2111–2140 (2002)
19. Farhat, C., Géradin, M.: On the general solution by a direct method of a large scale singular system of linear equations: application to the analysis of floating structures. Int. J. Numer. Methods Eng. **41**, 675–696 (1998)
20. Kornhuber, R., Krause, R.: Adaptive multigrid methods for Signorini's problem in linear elasticity. Comput. Vis. Sci. **4**(1), 9–20 (2001)
21. Kornhuber, R., Krause, R., Sander, O., Deuflhard, P., Ertel, S.: A monotone multigrid solver for two body contact problems in biomechanics. Comput. Vis. Sci. **11**, 3–15 (2008)
22. Iontcheva, A.H., Vassilevski, P.S.: Monotone multigrid methods based on element agglomeration coarsening away from the contact boundary for the Signorini's problem. Numer. Linear Algebra Appl. **11**(2–3), 189–204 (2004)
23. Simo, J.C., Laursen, T.A.: An augmented Lagrangian treatment of contact problems involving friction. Comput. Struct. **42**, 97–116 (1992)
24. Glowinski, R., Le Tallec, P.: Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics. SIAM, Philadelphia (1989)
25. Conn, A.R., Gould, N.I.M., Toint, Ph.L.: LANCELOT: A FORTRAN Package for Large Scale Nonlinear Optimization (Release A). No. 17 in Springer Series in Computational Mathematics. Springer, New York (1992)
26. Dostál, Z., Friedlander, A., Santos, S.A.: Solution of contact problems of elasticity by FETI domain decomposition. Domain Decomposition Methods 10. Contemporary Mathematics, vol. 218, 82–93. AMS, Providence (1998)

27. Dostál, Z., Gomes, F.A.M., Santos, S.A.: Duality based domain decomposition with natural coarse space for variational inequalities. J. Comput. Appl. Math. **126**(1–2), 397–415 (2000)
28. Dostál, Z., Gomes, F.A.M., Santos, S.A.: Solution of contact problems by FETI domain decomposition with natural coarse space projection. Comput. Methods Appl. Mech. Eng. **190**(13–14), 1611–1627 (2000)
29. Dostál, Z., Horák, D., Kučera, R., Vondrák, V., Haslinger, J., Dobiáš, J., Pták, S.: FETI based algorithms for contact problems: scalability, large displacements and 3D Coulomb friction. Comput. Methods Appl. Mech. Eng. **194**(2–5), 395–409 (2005)
30. Dostál, Z., Kozubek, T., Vondrák, V., Brzobohatý, T., Markopoulos, A.: Scalable TFETI algorithm for the solution of multibody contact problems of elasticity. Int. J. Numer. Methods Eng. **82**(11), 1384–1405 (2010)
31. Dostál, Z., Schöberl, J.: Minimizing quadratic functions over non-negative cone with the rate of convergence and finite termination. Comput. Optim. Appl. **30**(1), 23–44 (2005)
32. Dostál, Z.: Inexact semi-monotonic augmented Lagrangians with optimal feasibility convergence for quadratic programming with simple bounds and equality constraints. SIAM J. Numer. Anal. **43**(1), 96–115 (2005)
33. Dostál, Z.: Optimal Quadratic Programming Algorithms, with Applications to Variational Inequalities, 1st edn. Springer, New York (2009)
34. Dostál, Z., Horák, D., Kučera, R.: Total FETI - an easier implementable variant of the FETI method for numerical solution of elliptic PDE. Commun. Numer. Methods Eng. **22**, 1155–1162 (2006)
35. Toselli, A., Widlund, O.B.: Domain Decomposition Methods – Algorithms and Theory. Springer Series on Computational Mathematics, vol. 34. Springer, Berlin (2005)
36. Jarošová, M., Klawonn, A., Rheinbach, O.: Projector preconditioning and transformation of basis in FETI-DP algorithms for contact problems. Math. Comput. Simul. **82**(10), 1894–1907 (2012)

Chapter 12
Contact Problems with Friction

Contact problems become much more complicated when a kind of friction is taken into account, even in the case that we restrict our attention to 3D problems of linear elasticity. The problems start with the formulation of friction laws, which are of phenomenological nature. The most popular friction law, the Coulomb law of friction, makes the problem intrinsically non-convex. The other new difficulties which we face in the development of scalable algorithms for contact with friction include the non-uniqueness of a solution, nonlinear inequality constraints in dual formulation, and intractable minimization formulation.

To get some kind of useful optimality results, we concentrate our effort on the solution of a simpler problem with a given (Tresca) friction. This model of friction assumes that the normal pressure is a priori known on the contact interface. Though such assumption is not realistic and violates the laws of physics, e.g., it admits positive contact pressure on a part of contact interface that is not active in the solution, it makes the problem well-posed and its solution can be used in the fixed-point iterations for the solution of problems with Coulomb's friction. Though such algorithm is not supported by strong convergence theory, it can often find a solution of contact problems with friction in a small number of outer iterations.

It is important for our development that the solution of problems with Tresca's friction minimizes a convex nonsmooth cost function subject to a convex set. Switching to the dual formulation results in a convex QCQP (Quadratically Constrained Quadratic Programme) problem with linear non-penetration constrains and separable quadratic inequality constraints defined by the slip bounds. The basic structure of the problem complies well with TFETI. Rather surprisingly, the optimality results presented here are very similar to those proved for the frictionless problems.

The TFETI-based algorithms presented here can use tens of thousands of cores to solve effectively both coercive and semicoercive problems decomposed into tens of thousands of subdomains and billions of nodal variables. For larger problems, the initialization of the iterative solving procedure starts to dominate the costs. Some modifications for emerging exascale technologies are discussed in Chap. 19.

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3_12

12.1 Equilibrium of Bodies in Contact with Coulomb Friction

Let us consider a system of bodies $\Omega^1, \dots, \Omega^s$ introduced in Sects. 11.1 and 11.2, which is illustrated in Fig. 11.1, and let us examine the effect of friction on the equilibrium in the tangential plane of a smooth contact interface under the assumption that it complies with Coulomb's law.

Let us start by assuming that the contact coupling set $\mathscr{S}$ was specified and let $(p, q) \in \mathscr{S}$. Using the notation introduced in Sect. 11.2 and assuming that the displacement

$$\mathbf{u} = (\mathbf{u}^1, \dots, \mathbf{u}^s), \quad \mathbf{u}^p : \overline{\Omega}^p \to \mathbb{R}^3,$$

is sufficiently smooth, we can write the non-penetration contact conditions and the conditions of equilibrium in the normal direction $\mathbf{n}$ briefly as

$$[u_n] \le g, \quad \lambda_n \ge 0, \quad \lambda_n([u_n] - g) = 0, \quad \mathbf{x} \in \Gamma^{pq}. \tag{12.1}$$

Recall that λ_n denotes the normal traction on the slave face of the contact interface and $[u_n]$ denotes the jump of the boundary displacements. To formulate the conditions of equilibrium in the tangential plane, let us introduce the notation

$$\boldsymbol{\lambda}_T = \boldsymbol{\lambda} - \boldsymbol{\lambda}_N, \quad \mathbf{u}_N = u_n \mathbf{n}^p, \quad \mathbf{u}_T = \mathbf{u} - \mathbf{u}_N, \quad \mathbf{x} \in \Gamma_C^p, \quad p = 1, \dots, s,$$

and

$$[\mathbf{u}_T] = \mathbf{u}_T^p - \mathbf{u}_T^q \circ \chi^{pq}, \quad \mathbf{x} \in \Gamma_C^{pq}, \quad (p, q) \in \mathscr{S}.$$

Let $\Phi > 0$ denote the friction coefficient, which can depend on $\mathbf{x}$ provided it is bounded away from zero. The Coulomb friction law at $\mathbf{x} \in \Gamma_C^{pq}$, $(p, q) \in \mathscr{S}$, can be written in the form

$$\text{if } [u_n] = g, \quad \text{then} \quad \lambda_n \ge 0, \tag{12.2}$$

$$\text{if } [u_n] = g \text{ and } \|\boldsymbol{\lambda}_T\| < \Phi \lambda_n, \quad \text{then} \quad \mathbf{u}_T = \mathbf{o}, \tag{12.3}$$

$$\text{if } \|\boldsymbol{\lambda}_T\| = \Phi \lambda_n, \text{ then there is } \mu > 0 \text{ such that } [\mathbf{u}_T] = \mu [\boldsymbol{\lambda}_T], \tag{12.4}$$

where

$$g = \big(\chi - \mathsf{Id}\big) \cdot \mathbf{n}^p.$$

The relations (12.1)–(12.4) with constitutive relations (11.4) and

$$\begin{aligned} -\operatorname{div} \sigma(\mathbf{u}) &= \mathbf{f} \quad \text{in } \Omega, \\ \mathbf{u}^p &= \mathbf{o} \quad \text{on } \Gamma_U^p, \\ \sigma(\mathbf{u}^p)\mathbf{n}^p &= \mathbf{f}_\Gamma^p \text{ on } \Gamma_F, \end{aligned} \tag{12.5}$$

$p = 1, \dots, s$, describe completely the kinematics and equilibrium of a system of elastic bodies in contact obeying the Coulomb friction law.

12.2 Variational Formulation

The classical formulation of contact problems with friction (12.1)–(12.5) makes sense only when the solutions satisfy strong regularity assumptions similar to those required by the classical solutions of frictionless problems. The remedy is again a variational formulation which characterizes the solutions by local averages. We shall use the same tools as in Sect. 11.3, namely the spaces

$$V^p = \left\{ \mathbf{v} \in \left(H^1(\Omega^p)\right)^3 : \mathbf{v} = \mathbf{o} \text{ on } \Gamma_U^p \right\}, \ p = 1, \dots, s,$$
$$V = V^1 \times \cdots \times V^s,$$

the convex set

$$\mathscr{K} = \left\{ \mathbf{v} \in V : [v_n] \le g \text{ on } \Gamma_C^{pq}, \ (p,q) \in \mathscr{S} \right\},$$

and the forms

$$a(\mathbf{u}, \mathbf{v}) = \sum_{p=1}^{s} \int_{\Omega_p} \sigma\left(\mathbf{v}^p\right) : \varepsilon\left(\mathbf{u}^p\right) \, \mathrm{d}\Omega, \quad \mathbf{u}^p, \mathbf{v}^p \in V^p, \tag{12.6}$$

$$\ell(\mathbf{v}) = \sum_{p=1}^{s} \int_{\Gamma_F^p} \mathbf{f}_\Gamma^p \cdot (\mathbf{v}^p - \mathbf{u}^p) \, \mathrm{d}\Gamma + \int_{\Omega^p} \mathbf{f}^p \cdot (\mathbf{v}^p - \mathbf{u}^p) \, \mathrm{d}\Gamma, \quad \mathbf{u}^p, \mathbf{v}^p \in V^p \tag{12.7}$$

Moreover, we assume that a contact coupling set $\mathscr{S}$ has been chosen and we shall use the short notation of the previous section.

The effect of friction will be enhanced by a nonlinear functional j which represents the virtual work of frictional forces

$$j(\mathbf{u}, \mathbf{v}) = \sum_{(p,q) \in \mathscr{S}} \int_{\Gamma_C^{pq}} \Phi(|\lambda_n(\mathbf{u})|) \, \|[\mathbf{v}_T]\| \, \mathrm{d}\Gamma, \quad \mathbf{u}, \mathbf{v} \in V. \tag{12.8}$$

We shall start with identity (11.13) which reads

$$\begin{aligned} a(\mathbf{u}, \mathbf{v} - \mathbf{u}) - \ell(\mathbf{v} - \mathbf{u}) &= \sum_{p=1}^{s} \int_{\Gamma_C^p} \sigma(\mathbf{u}^p)\mathbf{n}^p \cdot (\mathbf{v}^p - \mathbf{u}^p) \, \mathrm{d}\Gamma \\ &= \sum_{(p,q) \in \mathscr{S}} \int_{\Gamma_C^{pq}} \boldsymbol{\lambda} \cdot (\mathbf{u} - \mathbf{v} + (\mathbf{v} - \mathbf{u}) \circ \chi) \, \mathrm{d}\Gamma \\ &= \sum_{(p,q) \in \mathscr{S}} \int_{\Gamma_C^{pq}} \boldsymbol{\lambda} \cdot ([\mathbf{u}] - [\mathbf{v}]) \, \mathrm{d}\Gamma, \quad \mathbf{u}, \mathbf{v} \in V, \end{aligned}$$

and show that the classical solution $\mathbf{u} \in \mathscr{K}$ of the contact problem with Coulomb's friction (12.1)–(12.5) satisfies

$$a(\mathbf{u}, \mathbf{v}-\mathbf{u}) - \ell(\mathbf{v}-\mathbf{u}) + j(\mathbf{u}, \mathbf{v}) - j(\mathbf{u}, \mathbf{u}) \geq 0, \quad \mathbf{v} \in \mathscr{K}. \tag{12.9}$$

First notice that using (12.9) and $\mathbf{v} \in \mathscr{K}$, we get

$$\begin{aligned}\boldsymbol{\lambda}(\mathbf{u}) \cdot ([\mathbf{u}] - [\mathbf{v}]) &= \boldsymbol{\lambda}_T(\mathbf{u}) \cdot ([\mathbf{u}_T] - [\mathbf{v}_T]) + \lambda_n(\mathbf{u}) \cdot ([u_n] - [v_n]) \\ &= \boldsymbol{\lambda}_T(\mathbf{u}) \cdot ([\mathbf{u}_T] - [\mathbf{v}_T]) + \lambda_n(\mathbf{u}) \cdot ([u_n] - g + g - [v_n]) \\ &\geq \boldsymbol{\lambda}_T(\mathbf{u}) \cdot ([\mathbf{u}_T] - [\mathbf{v}_T]).\end{aligned}$$

After rewriting the left-hand side of (12.9) and using the above inequality, we get

$$\begin{aligned}&a(\mathbf{u}, \mathbf{v}-\mathbf{u}) - \ell(\mathbf{v}-\mathbf{u}) + j(\mathbf{u}, \mathbf{v}) - j(\mathbf{u}, \mathbf{u}) \\ &\quad = \sum_{(p,q)\in\mathscr{S}} \int_{\Gamma_C^{pq}} \left(\boldsymbol{\lambda}(\mathbf{u}) \cdot ([\mathbf{u}] - [\mathbf{v}]) + \Phi|\lambda_n(\mathbf{u})| \, \|[\mathbf{v}_T]\| - \Phi|\lambda_n(\mathbf{u})| \, \|[\mathbf{u}_T]\|\right) \mathrm{d}\Gamma \\ &\quad \geq \sum_{(p,q)\in\mathscr{S}} \int_{\Gamma_C^{pq}} \left(\boldsymbol{\lambda}_T(\mathbf{u}) \cdot ([\mathbf{u}_T] - [\mathbf{v}_T]) + \Phi|\lambda_n(\mathbf{u})| \, (\|\mathbf{v}_T\| - \|\mathbf{u}_T\|)\right) \mathrm{d}\Gamma.\end{aligned}$$

To show that the last expression is greater or equal to zero, let $(p, q) \in \mathscr{S}$ be arbitrary but fixed, let

$$\Psi = \Phi|\lambda_n(\mathbf{u})|, \quad \mathbf{x} \in \Gamma_C^{pq},$$

denote the *slip bound*, and consider two cases.

If $\|\boldsymbol{\lambda}_T(\mathbf{u})\| < \Psi$, then $[\mathbf{u}_T] = \mathbf{o}$ and

$$\boldsymbol{\lambda}_T(\mathbf{u}) \cdot ([\mathbf{u}_T] - [\mathbf{v}_T]) + \Phi|\lambda_n(\mathbf{u})| \, (\|\mathbf{v}_T\| - \|\mathbf{u}_T\|) = \Psi\|[\mathbf{v}_T]\| - \boldsymbol{\lambda}_T(\mathbf{u}) \cdot [\mathbf{v}_T] \geq 0.$$

If $\|\boldsymbol{\lambda}_T(\mathbf{u})\| = \Psi$, then by (12.4) $[\mathbf{u}_T] = \mu[\boldsymbol{\lambda}_T(\mathbf{u})], \quad \mu \geq 0$, so

$$[\mathbf{u}_T] \cdot [\boldsymbol{\lambda}_T(\mathbf{u})] = \|[\mathbf{u}_T]\|\|[\boldsymbol{\lambda}_T(\mathbf{u})]\|$$

and for $\mathbf{x} \in \Gamma_C^{pq}$

$$\begin{aligned}&\boldsymbol{\lambda}_T(\mathbf{u}) \cdot ([\mathbf{u}_T] - [\mathbf{v}_T]) + \Phi|\lambda_n(\mathbf{u})| \, (\|[\mathbf{v}_T]\| - \|[\mathbf{u}_T]\|) \\ &\qquad \geq \Psi(\|[\mathbf{v}_T]\| - \|[\mathbf{u}_T]\|) + \boldsymbol{\lambda}_T(\mathbf{u}) \cdot [\mathbf{u}_T] - \Psi\|[\mathbf{v}_T]\| \\ &\qquad = \boldsymbol{\lambda}_T(\mathbf{u}) \cdot [\mathbf{u}_T] - \Psi\|[\mathbf{u}_T]\| \\ &\qquad = \|[\mathbf{u}_T]\|(\|[\boldsymbol{\lambda}_T(\mathbf{u})]\| - \Psi) = 0.\end{aligned}$$

We have thus proved that each classical solution $\mathbf{u}$ satisfies (12.9). Choosing special $\mathbf{v} \in \mathscr{K}$, it is also possible to show that each sufficiently smooth solution $\mathbf{u}$ of (12.9) is a classical solution of (12.1)–(12.5).

Though (12.9) is well defined for more general functions than the ones considered above, its formulation in more general spaces faces considerable difficulties. For example, such formulation requires that the trace of $\sigma(\mathbf{x})$ on Γ_C is well defined and the theoretical results concerning the existence and uniqueness require additional assumptions. Moreover, no true potential energy functional exists for the problem with friction. For more detailed discussion, see the books by Hlaváček et al. [1], Kikuchi and Oden [2], Eck, Jarůšek, and Krbec [3], and the comprehensive papers by Haslinger, Hlaváček, and Nečas [4] or Wohlmuth [5].

12.3 Tresca (Given) Isotropic Friction

We shall now consider a simplification of the variational inequality (12.9) that can be reduced to a minimization problem which can be solved effectively by the methods similar to those discussed in Chap. 11.

The basic idea is to assume that the normal traction λ_n on the contact interface is known. Though it is possible to contrive a physical situation which can be captured by such model, the main motivation is the possibility to use the simplified model in the fixed-point iterations for solving contact problems with Coulomb's friction. Using the assumptions, we can formulate the problem to find a sufficiently smooth displacement field

$$\mathbf{u} = (\mathbf{u}^1, \dots, \mathbf{u}^s), \quad \mathbf{u}^p : \overline{\Omega}^p \to \mathbb{R}^3,$$

such that for any $(p, q) \in \mathscr{S}$ and $\mathbf{x} \in \Gamma^{pq}$

$$\text{if } [u_n] = g, \quad \text{then} \quad \lambda_n \geq 0, \tag{12.10}$$

$$\text{if } [u_n] = g \quad \text{and} \quad \|\boldsymbol{\lambda}_T\| < \Psi, \quad \text{then} \quad \mathbf{u}_T = \mathbf{o}, \tag{12.11}$$

$$\text{if } \|\boldsymbol{\lambda}_T\| = \Psi, \quad \text{then there is } \mu > 0 \quad \text{such that} \quad [\mathbf{u}_T] = \mu[\boldsymbol{\lambda}_T], \tag{12.12}$$

where $\Psi \geq 0$ is a sufficiently smooth prescribed slip bound. After substituting into the friction functional j (12.8) and simplifying the notation, we get

$$j(\mathbf{v}) = \sum_{p,q \in \mathscr{S}} \int_{\Gamma_C^{pq}} \Psi \|[\mathbf{v}_T]\| \, \mathrm{d}\Gamma \tag{12.13}$$

and the variational problem (12.9) reduces to the problem to find $\mathbf{u} \in \mathscr{K}$ such that

$$a(\mathbf{u}, \mathbf{v} - \mathbf{u}) - \ell(\mathbf{v} - \mathbf{u}) + j(\mathbf{v}) - j(\mathbf{u}) \geq 0, \quad \mathbf{v} \in \mathscr{K}. \tag{12.14}$$

Observing that

$$\Psi(\mathbf{x}) \|[\mathbf{v}_T(\mathbf{x})]\| = \max_{\|\boldsymbol{\tau}\| \leq \Psi(\mathbf{x})} \boldsymbol{\tau} \cdot [\mathbf{v}_T(\mathbf{x})],$$

we can write the non-differentiable term in (12.13) as

$$j(\mathbf{v}) = \sum_{(p,q)\in\mathscr{S}} \int_{\Gamma_C^{pq}} \Psi \|[\mathbf{v}_T]\| \,\mathrm{d}\Gamma = \sum_{(p,q)\in\mathscr{S}} \int_{\Gamma_C^{pq}} \max_{\|\boldsymbol{\tau}\|\le\Psi(\mathbf{x})} \boldsymbol{\tau}\cdot[\mathbf{v}_T(\mathbf{x})] \,\mathrm{d}\Gamma. \tag{12.15}$$

For the development of scalable algorithms for solving contact problems with Tresca friction, it is important that any solution of the variational boundary value inequality (12.14) solves

$$\min f(\mathbf{v}) \quad \text{subject to } \mathbf{v}\in\mathscr{K}, \quad f(\mathbf{v}) = \frac{1}{2}a(\mathbf{v},\mathbf{v}) - \ell(\mathbf{v}) + j(\mathbf{v}), \tag{12.16}$$

where a and ℓ are defined in (12.6) and (12.7), respectively. To verify this claim, notice that if $\mathbf{u}, \mathbf{v}\in\mathscr{K}$ satisfy (12.14), then

$$\begin{aligned} f(\mathbf{v}) - f(\mathbf{u}) &= \frac{1}{2}a(\mathbf{v},\mathbf{v}) - \ell(\mathbf{v}) + j(\mathbf{v}) - \frac{1}{2}a(\mathbf{u},\mathbf{u}) + \ell(\mathbf{u}) - j(\mathbf{u}) \\ &= \frac{1}{2}a(\mathbf{v}+\mathbf{u}, \mathbf{v}-\mathbf{u}) - \ell(\mathbf{v}-\mathbf{u}) + j(\mathbf{v}) - j(\mathbf{u}) \\ &= a(\mathbf{u}, \mathbf{v}-\mathbf{u}) - \ell(\mathbf{v}-\mathbf{u}) + j(\mathbf{v}) - j(\mathbf{u}) + \frac{1}{2}\big(a(\mathbf{v}, \mathbf{v}-\mathbf{u}) - a(\mathbf{u}, \mathbf{v}-\mathbf{u})\big) \\ &\ge \frac{1}{2}a(\mathbf{v}-\mathbf{u}, \mathbf{v}-\mathbf{u}) \ge 0. \end{aligned}$$

We conclude that any classical solution of the contact problem with given (Tresca) friction (12.1), (12.10)–(12.12), and (12.5) satisfies

$$f(\mathbf{u}) \le f(\mathbf{v}), \quad \mathbf{v}\in\mathscr{K}. \tag{12.17}$$

Denoting by $\mathbf{u}(\Psi)$ a solution of (12.17), we can try to get a solution of the problem with Coulomb's friction by the fixed-point iterations

$$\begin{aligned} \mathbf{u}^0 &= \mathbf{u}(\Psi^0), \\ \mathbf{u}^{k+1} &= \mathbf{u}\left(\Phi\lambda_n(\mathbf{u}^k)\right). \end{aligned}$$

The procedure was proposed by Panagiotopoulos [6].

12.4 Orthotropic Friction

Let us briefly describe *orthotropic Coulomb friction*, which is defined by the matrix

$$\Phi = \begin{bmatrix} \Phi_1 & 0 \\ 0 & \Phi_2 \end{bmatrix}, \quad \Phi_1, \Phi_2 > 0.$$

The diagonal entries of Φ describe the friction in two orthogonal tangential directions which are defined at each point of the contact interface by the unit vectors $\mathbf{t}_1, \mathbf{t}_2$. The friction coefficients Φ_1, Φ_2 can depend on $\mathbf{x}$, but in this case we assume that they are bounded away from zero. The orthotropic Coulomb friction law at $\mathbf{x} \in \Gamma_C^{pq}$, $(p, q) \in \mathscr{S}$, reads

$$\text{if } [u_n] = g, \quad \text{then} \quad \lambda_n \geq 0, \tag{12.18}$$

$$\text{if } [u_n] = g \text{ and } \|\Phi^{-1}\boldsymbol{\lambda}_T\| < \lambda_n, \quad \text{then} \quad \mathbf{u}_T = \mathbf{o}, \tag{12.19}$$

$$\text{if } |\Phi^{-1}\boldsymbol{\lambda}_T\| = \lambda_n, \text{ then there is } \mu > 0 \text{ such that } [\mathbf{u}_T] = \mu[\Phi^{-1}\boldsymbol{\lambda}_T]. \tag{12.20}$$

The variational formulation of the conditions of equilibrium for the orthotropic Coulomb friction is formally identical to (12.9) provided we use the dissipative term in the form

$$j(\mathbf{u}, \mathbf{v}) = \sum_{(p,q)\in\mathscr{S}} \int_{\Gamma_C^{pq}} |\lambda_n(\mathbf{u})| \, \|[\Phi\mathbf{v}_T]\| \, \mathrm{d}\Gamma, \quad \mathbf{u}, \mathbf{v} \in V.$$

The same is true for the variational formulation of the conditions of equilibrium for orthotropic Tresca friction, which is formally identical to (12.14) and (12.1) provided we use the non-differentiable term in the form

$$j(\mathbf{v}) = \sum_{(p,q)\in\mathscr{S}} \int_{\Gamma_C^{pq}} \Lambda_n \|[\Phi\mathbf{v}_T]\| \, \mathrm{d}\Gamma,$$

where $\Lambda_n \geq 0$ now defines the predefined normal traction. The reasoning is the same as in Sect. 12.3. Observing that

$$\begin{aligned} \Lambda_n \|[\Phi\mathbf{v}_T]\| &= \max_{\|\boldsymbol{\tau}\| \leq \Lambda_n} \boldsymbol{\tau} \cdot [\Phi\mathbf{v}_T] = \max_{\|\boldsymbol{\tau}\| \leq \Lambda_n} \Phi\boldsymbol{\tau} \cdot [\mathbf{v}_T] \\ &= \max_{\|\Phi^{-1}\Phi\boldsymbol{\tau}\| \leq \Lambda_n} \Phi\boldsymbol{\tau} \cdot [\mathbf{v}_T] = \max_{\|\Phi^{-1}\boldsymbol{\tau}\| \leq \Lambda_n} \boldsymbol{\tau} \cdot [\mathbf{v}_T], \end{aligned}$$

we can replace the non-differentiable term in (12.13) by the bilinear term to get

$$j(\mathbf{v}) = \sum_{(p,q)\in\mathscr{S}} \int_{\Gamma_C^{pq}} \Lambda_n \|[\Phi\mathbf{v}_T]\| \, \mathrm{d}\Gamma = \sum_{(p,q)\in\mathscr{S}} \int_{\Gamma_C^{pq}} \max_{\|\Phi^{-1}\boldsymbol{\tau}\| \leq \Lambda_n} \boldsymbol{\tau} \cdot [\mathbf{v}_T(\mathbf{x})] \, \mathrm{d}\Gamma. \tag{12.21}$$

Denoting by $\mathbf{u}(\Lambda_n)$ a solution of (12.17) with j defined in (12.21), we can try to get a solution of the problem with orthotropic friction by the fixed-point iterations

$$\begin{aligned} \mathbf{u}^0 &= \mathbf{u}(\Lambda_n^0), \\ \mathbf{u}^{k+1} &= \mathbf{u}\left(\lambda_n(\mathbf{u}^k)\right). \end{aligned}$$

To simplify the exposition, we shall restrict our attention in what follows to the isotropic friction, leaving the modification to the interested reader.

12.5 Domain Decomposition and Discretization

The problem (12.17) has the same feasible set as the frictionless problem (11.15) and differs only in the cost function. It turns out that we can use the domain decomposition in the same way as in Sect. 11.4. To this end, let us decompose each domain Ω^p into subdomains with sufficiently smooth boundaries, assign a unique number to each subdomain, decompose appropriately the parts of the boundaries, introduce their numbering as in Sect. 11.4, and introduce the notation

$$
\begin{aligned}
V_{DD}^p &= \left\{ \mathbf{v} \in H^1(\Omega^p)^3 : \mathbf{v} = \mathbf{o} \ \text{ on } \ \Gamma_U^p \right\}, \quad p = 1, \dots s, \\
V_{DD} &= V^1 \times \cdots \times V^s, \\
\mathcal{K}_{DD} &= \left\{ \mathbf{v} \in V_{DD} : [v_n] \le g \text{ on } \Gamma_C^{pq}, \ (p,q) \in \mathscr{S}; \ \mathbf{v}^p = \mathbf{v}^q \text{ on } \Gamma_G^{pq}, \ p, q = 1, \dots, s \right\}.
\end{aligned}
$$

The variational decomposed problem with the Tresca (given) friction defined by a slip bound Ψ reads

$$
\text{find } \ \mathbf{u} \in \mathcal{K}_{DD} \ \text{ such that } \ f(\mathbf{u}) \le f(\mathbf{v}), \quad \mathbf{v} \in \mathcal{K}_{DD}, \tag{12.22}
$$

where

$$
f(\mathbf{v}) = \frac{1}{2} a(\mathbf{v}, \mathbf{v}) - \ell(\mathbf{v}) + j(\mathbf{v}).
$$

To get the discretized problem (12.22), let us use a quasi-uniform discretization with shape regular elements as in Sect. 11.5 and apply the procedure described in Sect. 11.5 to the components of the discretized linear form ℓ, the quadratic form a, the gap function g, and the matrices that describe the discretized feasible set $\mathcal{K}_{DD}$. Thus we shall get the block diagonal stiffness matrix $\mathsf{K} \in \mathbb{R}^{n \times n}$, the vector of nodal forces $\mathbf{f} \in \mathbb{R}^n$, the constraint matrix $\mathsf{B}_I \in \mathbb{R}^{m_I}$, the discretized gap function $\mathbf{c}_I \in \mathbb{R}^{m_I}$, and the matrix $\mathsf{B}_E \in \mathbb{R}^{m_I}$ that describes the gluing and Dirichlet conditions.

To define the discretized non-differentiable term j (12.13), we introduce $2m_I \times n$ matrix T such that m_I blocks of which are $\mathsf{T}_i = \mathsf{T}_i(\mathbf{x}_i) \in \mathbb{R}^{2 \times n}$ are formed by appropriately placed orthonormal tangential vectors $\mathbf{t}_1(\mathbf{x}_i)$ and $\mathbf{t}_2(\mathbf{x}_i)$ so that the tangential component of the displacement $\mathbf{u}$ is given by $\mathsf{T}_i \mathbf{u}$. After applying numerical integration to (12.13), we get the discretized dissipative term in the form

$$
j(\mathbf{u}) = \sum_{i=1}^{m_c} \Psi_i \|\mathsf{T}_i \mathbf{u}\|, \tag{12.23}
$$

where Ψ_i is the slip bound associated with $\mathbf{x}_i$. Using (12.15), we get

$$j(\mathbf{u}) = \sum_{i=1}^{m_c} \Psi_i \|\mathsf{T}_i \mathbf{u}\| = \sum_{i=1}^{m_c} \max_{\|\boldsymbol{\tau}_i\| \le \Psi_i} \boldsymbol{\tau}_i^T \mathsf{T}_i \mathbf{u}, \tag{12.24}$$

where $\boldsymbol{\tau}_i \in \mathbb{R}^2$ can be interpreted as a Lagrange multiplier. The right-hand side of (12.24) is differentiable.

The discretized problem (12.22) reads

$$\text{find } \mathbf{u} \in \mathscr{K} \text{ such that } f(\mathbf{u}) \le f(\mathbf{v}), \quad \mathbf{v} \in \mathscr{K}, \tag{12.25}$$

where

$$f(\mathbf{v}) = \frac{1}{2}\mathbf{v}^T \mathsf{K}\mathbf{v} - \mathbf{f}^T \mathbf{v} + j(\mathbf{v}), \quad \mathscr{K} = \{\mathbf{v} \in \mathbb{R}^n : \mathsf{B}_I \mathbf{u} \le \mathbf{c}_I \text{ and } \mathsf{B}_E \mathbf{u} = \mathbf{o}\}. \tag{12.26}$$

We assume that B_I, B_E, and T are full rank matrices. Notice that

$$\mathsf{T}^T \mathsf{B}_I = \mathsf{O}.$$

If each node is involved in at most one inequality, then the rows of B_I can be orthogonal and it is possible (see Remark 11.1) to achieve that

$$\mathsf{B}\mathsf{B}^T = \mathsf{I}, \quad \mathsf{B} = \begin{bmatrix} \mathsf{B}_E \\ \mathsf{B}_I \\ \mathsf{T} \end{bmatrix}. \tag{12.27}$$

12.6 Dual Formulation

The problem (12.25) is not suitable for numerical solution, even if we replace j in the cost function by (12.24). The reasons are that the stiffness matrix K is typically large, ill-conditioned, and singular and the feasible set is in general so complex that the projections onto it can hardly be effectively computed. Under the circumstances, it would be very difficult to solve efficiently auxiliary problems and to effectively identify the solution active set.

As in Chap. 11, the complications can be essentially reduced by applying the duality theory of convex programming (see, e.g., Bazaraa, Shetty, and Sherali [7]). In the dual formulation of problem (12.25), we use three types of Lagrange multipliers, namely $\boldsymbol{\lambda}_I \in \mathbb{R}^{m_I}$ associated with the non-interpenetration condition, $\boldsymbol{\lambda}_E \in \mathbb{R}^{m_E}$ associated with the "gluing" and prescribed displacements, and

$$\boldsymbol{\tau} = [\boldsymbol{\tau}_1^T, \boldsymbol{\tau}_2^T, \ldots, \boldsymbol{\tau}_{m_T}^T]^T \in \mathbb{R}^{2m_T}, \quad \boldsymbol{\tau}_i \in \mathbb{R}^2, \quad i = 1, \ldots, m_T,$$

which are used to smooth the non-differentiability. The Lagrangian associated with problem (12.25) reads

$$L(\mathbf{u}, \boldsymbol{\lambda}_I, \boldsymbol{\lambda}_E, \boldsymbol{\tau}) = f(\mathbf{u}) + \boldsymbol{\tau}^T \mathsf{T}\mathbf{u} + \boldsymbol{\lambda}_I^T(\mathsf{B}_I\mathbf{u} - \mathbf{c}_I) + \boldsymbol{\lambda}_E^T(\mathsf{B}_E\mathbf{u} - \mathbf{c}_E).$$

To simplify the notation, we denote $m = m_E + m_I + 2m_T = m_E + 3m_I$,

$$\boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{\lambda}_E \\ \boldsymbol{\lambda}_I \\ \boldsymbol{\tau} \end{bmatrix}, \quad \mathsf{B} = \begin{bmatrix} \mathsf{B}_E \\ \mathsf{B}_I \\ \mathsf{T} \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} \mathbf{o} \\ \mathbf{c}_I \\ \mathbf{o} \end{bmatrix},$$

and

$$\Lambda(\Psi) = \{(\boldsymbol{\lambda}_E^T, \boldsymbol{\lambda}_I^T, \boldsymbol{\tau}^T)^T \in \mathbb{R}^m : \boldsymbol{\lambda}_I \geq \mathbf{o}, \|\tau_i\| \leq \Psi_i, \ i = 1, \ldots, m_T\},$$

so that we can write the Lagrangian briefly as

$$L(\mathbf{u}, \boldsymbol{\lambda}) = \frac{1}{2}\mathbf{u}^T \mathsf{K}\mathbf{u} - \mathbf{f}^T\mathbf{u} + \boldsymbol{\lambda}^T(\mathsf{B}\mathbf{u} - \mathbf{c}).$$

Using the convexity of the cost function and constraints, we can reformulate problem (12.25) by duality to get

$$\min_{\mathbf{u}} \sup_{\boldsymbol{\lambda} \in \Lambda(\Psi)} L(\mathbf{u}, \boldsymbol{\lambda}_I, \boldsymbol{\lambda}_E, \boldsymbol{\tau}) = \max_{\boldsymbol{\lambda} \in \Lambda(\Psi)} \min_{\mathbf{u}} L(\mathbf{u}, \boldsymbol{\lambda}_I, \boldsymbol{\lambda}_E, \boldsymbol{\tau}).$$

We conclude that problem (12.25) is equivalent to the saddle point problem

$$L(\widehat{\mathbf{u}}, \widehat{\boldsymbol{\lambda}}) = \max_{\boldsymbol{\lambda} \in \Lambda(\Psi)} \min_{\mathbf{u}} L(\mathbf{u}, \boldsymbol{\lambda}). \tag{12.28}$$

Recall that B is a full rank matrix. For a fixed $\boldsymbol{\lambda}$, the Lagrange function $L(\cdot, \boldsymbol{\lambda})$ is convex in the first variable and the minimizer $\mathbf{u}$ of $L(\cdot, \boldsymbol{\lambda})$ satisfies

$$\mathsf{K}\mathbf{u} - \mathbf{f} + \mathsf{B}^T\boldsymbol{\lambda} = \mathbf{o}. \tag{12.29}$$

Equation (12.29) has a solution if and only if

$$\mathbf{f} - \mathsf{B}^T\boldsymbol{\lambda} \in \mathrm{Im}\mathsf{K}, \tag{12.30}$$

which can be expressed more conveniently by means of a matrix $\mathsf{R} \in \mathbb{R}^{n \times 6s}$ the columns of which span the null space of K as

$$\mathsf{R}^T(\mathbf{f} - \mathsf{B}^T\boldsymbol{\lambda}) = \mathbf{o}.$$

The matrix R can be formed directly, using any basis of the rigid body modes of the subdomains. See Sect. 11.6 for details.

Now assume that $\boldsymbol{\lambda}$ satisfies (12.30) and denote by K^+ any matrix which satisfies

$$\mathsf{K}\mathsf{K}^+\mathsf{K} = \mathsf{K}. \tag{12.31}$$

Let us note that the action of a generalized inverse which satisfies (12.31) can be evaluated at the cost comparable with that of Cholesky's decomposition applied to the regularized K (see Sect. 11.7). It can be verified directly that if $\mathbf{u}$ solves (12.29), then there is a vector $\alpha \in \mathbb{R}^{6s}$ such that

$$\mathbf{u} = \mathsf{K}^+(\mathbf{f} - \mathsf{B}^T\boldsymbol{\lambda}) + \mathsf{R}\alpha. \tag{12.32}$$

After substituting expression (12.32) into problem (12.28), changing the signs, and omitting the constant term, we get that $\boldsymbol{\lambda}$ solves the minimization problem

$$\min \ \Theta(\boldsymbol{\lambda}) \quad \text{s.t.} \quad \boldsymbol{\lambda} \in \Lambda(\Psi) \quad \text{and} \quad \mathsf{R}^T(\mathbf{f} - \mathsf{B}^T\boldsymbol{\lambda}) = \mathbf{o}, \tag{12.33}$$

where

$$\Theta(\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\lambda}^T\mathsf{B}\mathsf{K}^+\mathsf{B}^T\boldsymbol{\lambda} - \boldsymbol{\lambda}^T(\mathsf{B}\mathsf{K}^+\mathbf{f} - \mathbf{c}).$$

Once the solution $\widehat{\boldsymbol{\lambda}}$ of (12.33) is known, the solution $\widehat{\mathbf{u}}$ of (12.25) can be evaluated by (12.32) with

$$\boldsymbol{\alpha} = (\mathsf{R}^T\widetilde{\mathsf{B}}^T\widetilde{\mathsf{B}}\mathsf{R})^{-1}\mathsf{R}^T\widetilde{\mathsf{B}}^T(\widetilde{\mathbf{c}} - \widetilde{\mathsf{B}}\mathsf{K}^+(\mathbf{f} - \mathsf{B}^T\widehat{\boldsymbol{\lambda}})),$$

where $\widetilde{\mathsf{B}}$ and $\widetilde{\mathbf{c}}$ are formed by the rows of B and $\mathbf{c}$ corresponding to all equality constraints and all free inequality constraints.

12.7 Preconditioning by Projectors to Rigid Body Modes

Even though problem (12.33) is much more suitable for computations than (12.25), further improvement can be achieved by adapting the observations that we used in the previous chapters. Let us denote

$$\begin{aligned} \widetilde{\mathsf{F}} &= \mathsf{B}\mathsf{K}^+\mathsf{B}^T, & F &= \|\widetilde{\mathsf{F}}\|, \\ \mathsf{F} &= F^{-1}\widetilde{\mathsf{F}}, & \widetilde{\mathbf{d}} &= F^{-1}(\mathsf{B}\mathsf{K}^+\mathbf{f} - \mathbf{c}), \\ \widetilde{\mathsf{G}} &= \mathsf{R}^T\mathsf{B}^T, & \widetilde{\mathbf{e}} &= \mathsf{R}^T\mathbf{f}, \end{aligned}$$

and let U denote a regular matrix that defines the orthonormalization of the rows of $\widetilde{\mathbf{G}}$ so that the matrix

$$\mathsf{G} = \mathsf{U}\widetilde{\mathsf{G}}$$

has orthonormal rows. After denoting

$$\mathbf{e} = \mathsf{U}\widetilde{\mathbf{e}},$$

problem (12.33) reads

$$\min \quad \frac{1}{2}\boldsymbol{\lambda}^T\mathsf{F}\boldsymbol{\lambda} - \boldsymbol{\lambda}^T\widetilde{\mathbf{d}} \quad \text{s.t.} \quad \boldsymbol{\lambda} \in \Lambda(\Psi) \quad \text{and} \quad \mathsf{G}\boldsymbol{\lambda} = \mathbf{e}. \tag{12.34}$$

Next we transform the problem of minimization on a subset of the affine space to that on a subset of the vector space using $\widetilde{\boldsymbol{\lambda}}$ which satisfies

$$\mathsf{G}\widetilde{\boldsymbol{\lambda}} = \mathbf{e}. \tag{12.35}$$

We can choose $\widetilde{\boldsymbol{\lambda}}$ similarly as we did in Sect. 11.8. In particular, if the problem is coercive, then we can use Lemma 11.1 to get

$$\widetilde{\boldsymbol{\lambda}} = \begin{bmatrix} \mathbf{o}_{\mathscr{I}} \\ \mathsf{G}_{\mathscr{E}}^T(\mathsf{G}_{\mathscr{E}}\mathsf{G}_{\mathscr{E}}^T)^{-1}\mathbf{e} \end{bmatrix},$$

or to use $\widetilde{\boldsymbol{\lambda}}$ which solves

$$\min \quad \frac{1}{2}\|\boldsymbol{\lambda}\|^2 \quad \text{s.t.} \quad \boldsymbol{\lambda} \in \Lambda(\Psi) \quad \text{and} \quad \mathsf{G}\boldsymbol{\lambda} = \mathbf{e}. \tag{12.36}$$

The above choices of $\widetilde{\boldsymbol{\lambda}}$ are important for the proof of optimality as they guarantee that $\mathbf{o}$ is a feasible vector of the modified problem. For practical computations, we can use the least squares solution of $\mathsf{G}\boldsymbol{\lambda} = \mathbf{e}$ given by

$$\widetilde{\boldsymbol{\lambda}} = \mathsf{G}^T\mathbf{e}.$$

Having $\widetilde{\boldsymbol{\lambda}}$, we can look for the solution of (12.34) in the form $\boldsymbol{\lambda} = \boldsymbol{\mu} + \widetilde{\boldsymbol{\lambda}}$. To carry out the transformation, denote $\boldsymbol{\lambda} = \boldsymbol{\mu} + \widetilde{\boldsymbol{\lambda}}$ and $\widetilde{\Lambda}(\Psi) = \Lambda(\Psi) - \widetilde{\boldsymbol{\lambda}}$, so that

$$\frac{1}{2}\boldsymbol{\lambda}^T\mathsf{F}\boldsymbol{\lambda} - \boldsymbol{\lambda}^T\widetilde{\mathbf{d}} = \frac{1}{2}\boldsymbol{\mu}^T\mathsf{F}\boldsymbol{\mu} - \boldsymbol{\mu}^T(\widetilde{\mathbf{d}} - \mathsf{F}\widetilde{\boldsymbol{\lambda}}) + \frac{1}{2}\widetilde{\boldsymbol{\lambda}}^T\mathbf{F}\widetilde{\boldsymbol{\lambda}} - \widetilde{\boldsymbol{\lambda}}^T\widetilde{\mathbf{d}}$$

and problem (12.34) turns, after returning to the old notation, into

$$\min \quad \frac{1}{2}\boldsymbol{\lambda}^T\mathsf{F}\boldsymbol{\lambda} - \boldsymbol{\lambda}^T\mathbf{d} \quad \text{s.t.} \quad \mathsf{G}\boldsymbol{\lambda} = \mathbf{o} \quad \text{and} \quad \boldsymbol{\lambda} \in \widetilde{\Lambda}(\Psi), \quad \mathbf{d} = \widetilde{\mathbf{d}} - \mathsf{F}\widetilde{\boldsymbol{\lambda}}. \tag{12.37}$$

Recall that we can achieve that $\mathbf{o} \in \widetilde{\Lambda}$.

Our final step is based on the observation that problem (12.37) is equivalent to

$$\min \quad \overline{\theta}_\rho(\boldsymbol{\lambda}) \quad \text{s.t.} \quad \mathsf{G}\boldsymbol{\lambda} = \mathbf{o} \quad \text{and} \quad \boldsymbol{\lambda} \in \widetilde{\Lambda}(\Psi), \tag{12.38}$$

where ρ is an arbitrary positive constant,

$$\mathsf{Q} = \mathsf{G}^T\mathsf{G} \qquad \text{and} \qquad \mathsf{P} = \mathsf{I} - \mathsf{Q}$$

denote the orthogonal projectors on the image space of G^T and on the kernel of G, respectively, and

$$\overline{\theta}_\rho(\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\lambda}^T(\mathsf{PFP} + \rho\mathsf{Q})\boldsymbol{\lambda} - \boldsymbol{\lambda}^T\mathsf{P}\mathbf{d}.$$

12.8 Optimality

To show that Algorithm 9.1 (SMALSE) with the inner loop implemented by Algorithm 7.2 (MPGP) is optimal for the solution of a class of problems arising from varying discretizations of a given contact problem with Tresca friction, let us introduce a notation which complies with that used in Part II.

Let $C \geq 2$ and $\rho > 0$ denote given constants and let

$$\mathscr{T}_C = \{(H, h) \in \mathbb{R}^2 : \quad H/h \leq C\}$$

denote the set of indices. For any $t \in \mathscr{T}_C$, let us define

$$\begin{array}{ll} \mathsf{A}_t = \mathsf{PFP} + \rho\mathsf{Q}, & \mathbf{b}_t = \mathsf{P}\mathbf{d}, \\ \mathsf{B}^t = \mathsf{G}, & \Omega_S^t = \widetilde{\Lambda}(\Psi), \end{array}$$

where the vectors and matrices are those arising from the discretization of (12.22) with the discretization and decomposition parameters H and h, $t = (H, h)$. We shall assume that the discretization satisfies the assumptions of Theorem 11.1, including (12.27), and that $\mathbf{o} \in \Omega_S^t$. We get a class of problems

$$\min \quad f_t(\boldsymbol{\lambda}_t) \quad \text{s.t.} \; \mathsf{B}_t\boldsymbol{\lambda}_t = \mathbf{o} \; \text{ and } \; \boldsymbol{\lambda}_t \in \Omega_S^t \tag{12.39}$$

with

$$f_t(\boldsymbol{\lambda}_t) = \frac{1}{2}\boldsymbol{\lambda}_t^T\mathsf{A}_t\boldsymbol{\lambda}_t - \mathbf{b}_t^T\boldsymbol{\lambda}_t.$$

To see that the class of problems (12.39) satisfies the assumptions of Theorem 9.4, recall that $\mathsf{G}\mathsf{G}^T = \mathsf{I}$, so

$$\|\mathsf{B}_t\| \leq 1. \tag{12.40}$$

It follows by Theorem 11.1 that there are constants $a^C_{\max} > a^C_{\min} > 0$ such that

$$a^C_{\min} \le \alpha_{\min}(\mathsf{A}_t) \le \alpha_{\max}(\mathsf{A}_t) \le a^C_{\max} \tag{12.41}$$

for any $t \in \mathscr{T}_C$. Thus the assumptions of Theorem 9.4 are satisfied for any set of indices $\mathscr{T}_C$, $C > 2$, and $\rho > 0$. Using the arguments specified in the above discussion, we can state our main result:

Theorem 12.1 *Let $C \ge 2$, $\rho > 0$, and $\varepsilon > 0$ denote given constants, let $\{\boldsymbol{\lambda}^k_t\}$, $\{\boldsymbol{\mu}^k_t\}$, and $\{M_{t,k}\}$ be generated by Algorithm 9.1 (SMALSE-M) for problem (12.39) with parameters*

$$\|\mathbf{b}_t\| \ge \eta_t > 0, \quad 0 < \beta < 1, \quad M_{t,0} = M_0 > 0, \quad \rho > 0, \quad \text{and} \quad \boldsymbol{\mu}^0_t = \mathbf{o}.$$

Let Step 1 of SMALSE-M be implemented by Algorithm 7.2 (MPGP) with the parameters

$$\Gamma > 0 \quad \text{and} \quad \alpha \in (0, 2/a^C_{\max}),$$

so that it generates the iterates

$$\boldsymbol{\lambda}^{k,0}_t, \boldsymbol{\lambda}^{k,1}_t, \ldots, \boldsymbol{\lambda}^{k,l}_t = \boldsymbol{\lambda}^k_t$$

for the solution of (12.39) starting from $\boldsymbol{\lambda}^{k,0}_t = \boldsymbol{\lambda}^{k-1}_t$ with $\boldsymbol{\lambda}^{-1}_t = \mathbf{o}$, where $l = l_{t,k}$ is the first index satisfying

$$\|\mathbf{g}^P(\boldsymbol{\lambda}^{k,l}_t, \boldsymbol{\mu}^k_t, \rho)\| \le M_{t,k}\|\mathsf{C}_t\boldsymbol{\lambda}^{k,l}_t\|$$

or

$$\|\mathbf{g}^P(\boldsymbol{\lambda}^{k,l}_t, \boldsymbol{\mu}^k_t, \rho)\| \le \varepsilon\|\mathbf{b}_t\| \quad \text{and} \quad \|\mathsf{C}_t\boldsymbol{\lambda}^{k,l}_t\| \le \varepsilon\|\mathbf{b}_t\|. \tag{12.42}$$

Then for any $t \in \mathscr{T}_C$ and problem (12.39), Algorithm 9.1 generates an approximate solution $\boldsymbol{\lambda}^{k_t}_t$ which satisfies (12.42) at $O(1)$ matrix–vector multiplications by the Hessian A_t of f_t.

12.9 Numerical Experiments

Here we give some results that illustrate the performance of the TFETI-based algorithms implemented for the solution of contact problems with friction using their implementation in `MatSol` [8]. All the computations were carried out with the parameters recommended in Chaps. 7–9, i.e., $M_0 = 1$, $\eta = 0.1\|\mathsf{P}\mathbf{d}\|$, $\rho = 1 \approx \|\mathsf{PFP}\|$, $\Gamma = 1$, $\beta = 0.2$, $\lambda^0 = \mathbf{o}$, and $\mu^0 = \mathbf{o}$. The relative precision of the computations was $\varepsilon = 10^{-4}$ (see (9.40)).

12.9.1 Academic Benchmark

To demonstrate the efficiency of TFETI for the solution of contact problems with friction, we consider the two cantilever beams benchmark described in Sect. 11.11. The geometry of the benchmark is depicted in Fig. 11.4. Its material properties are defined in Sect. 11.11.

To demonstrate the scalability, we resolved the problem in Fig. 11.4 with the Tresca friction defined by the friction coefficient $\Phi = 0.1$. The slip bounds were defined as the product of normal traction from the solution of the frictionless problem of Sect. 11.11 and Φ. The discretizations and the decompositions were defined by the discretization parameter h and the decomposition parameter H, respectively. For each h and H, the bodies were decomposed into cubic subdomains and discretized by the regular mesh with hexahedral elements. We kept $H/h = 8$, so that the assumptions of Theorem 12.1 were satisfied. The normal and tangential tractions along the axis of the contact interface are shown in Fig. 12.1. This figure also shows that in the solution of the largest problem, most of the linear and quadratic inequality constraints were active.

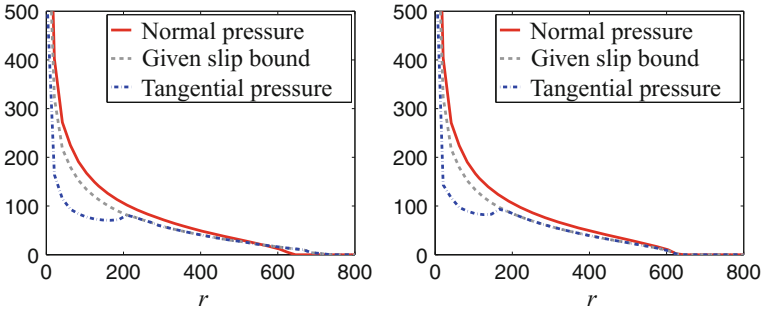

Fig. 12.1 Tresca (*left*) and Coulomb (*right*) friction – normal and contact pressures along r

The performance of the algorithms is documented in the following graphs. The numbers of outer and inner iterations of SMALSE and the multiplications by the Hessian F of the dual function for the problem with Tresca friction depending on the primal dimension n are depicted in Fig. 12.2. We can see stable numbers of both inner and outer iterations for n ranging from 431,244 to 11,643,588. The dual dimension of the problems ranged from 88,601 to 2,728,955. We conclude that the performance of the algorithm is in agreement with the theory. The problem is difficult because many nodes on the contact interface nearly touch each other, making the identification of the contact interface difficult.

Similar results were obtained for the solution of the problem with the Coulomb friction defined by the friction coefficient $\Phi = 0.1$. The results are in the above graphs. The number of iterations is naturally grater.

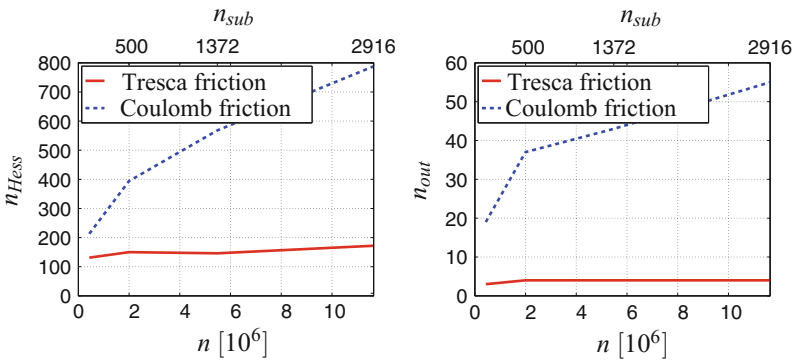

Fig. 12.2 Tresca and Coulomb: numerical scalability of cantilever beams–matrix–vector multiplications by F and outer iterations (*right*)

12.9.2 Yielding Clamp Connection

We have tested our algorithms on the solution of many real-world problems. Here we consider contact with the Coulomb friction, where the coefficient of friction was $\Phi = 0.5$. The problem was decomposed into 250 subdomains using METIS (see Fig. 12.3) and discretized by 1,592,853 and 216,604 primal and dual variables, respectively. The total displacements for the Coulomb friction are depicted in Fig. 12.4. It required 1,922 matrix–vector multiplications to find the solution.

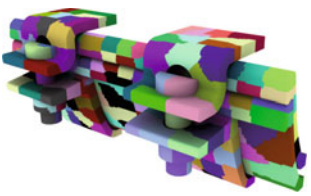

Fig. 12.3 Yielding clamp connection – domain decomposition

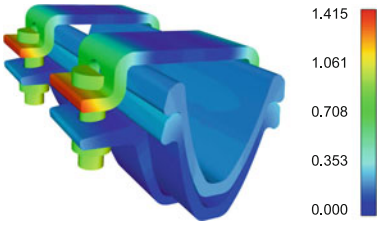

Fig. 12.4 Coulomb friction: total displacement

12.10 Comments and References

Early results concerning the formulation of contact problems with friction can be found in Duvaut and Lions [9]. An up-to-date overview of the theoretical results can be found in the book by Eck, Jarůšek, and Krbec [3]. See also Haslinger, Hlaváček, and Nečas [4]. A convenient presentation of the variational formulation, the results concerning the existence and uniqueness of a solution, the finite element approximation of the solution, and standard iterative methods for the solution of contact problems with friction can be found in the books by Kikuchi and Oden [2]. For useful theoretical results, see Hlaváček et al. [1] or Capatina [10]. A comprehensive presentation of the variationally consistent discretization of contact problems with friction using biorthogonal mortars (see also Chap. 15), including the approximation theory and solution algorithms based on the combination of the semi-smooth Newton method and multigrid, can be found in the seminal paper by Wohlmuth [5]. An up-to-date engineering approach to the solution of contact problems with friction can be found in Laursen [11] or Wriggers [12].

In spite of the difficulties associated with the formulation of contact problems with friction, a number of interesting results concerning the optimal complexity of algorithms for solving auxiliary problems were achieved by means of multigrid. The first papers adapted the earlier approach of Kornhuber and Krause to extend the results on frictionless problems to the problems with friction and gave an experimental evidence that it is possible to solve frictional two-body contact problems in 2D and 3D with multigrid efficiency (see, e.g., Wohlmuth and Krause [13], Krause [14], or Krause [15]).

The semi-smooth Newton method [16] became popular due to its capability to capture several nonlinearities in one loop (see, e.g., Hager and Wohlmuth [17], Wohlmuth [5], or comments in Sect. 17.6). The method is closely related to the active set strategy [18]. The problems suitable for the application of semi-smooth Newton include combination of contact with electrostatics (see, e.g., Migorski, Ochal, and Sofonea [19] or Hüeber, Matei, and Wohlmuth [20]). The lack of global convergence theory was compensated by some globalization strategies (see, e.g., Ito and Kunisch [21] or [22]).

The development of domain decomposition-based algorithms for contact problems with friction used from the early beginning the scheme proposed by Panagiotopoulos [6] which combines the outer fix point iterations for normal contact pressure with the solution of a problem with given friction in the inner loop (see Dostál and Vondrák [23] or Dostál et al. [24]). Though the convergence of the outer loop is supported by the theory only for very special cases [25], the method works well in practice. However, it seems that there is a little chance to develop the theory for contact problems with Coulomb friction as complete as that for the frictionless problems in the previous chapter. Thus the effort to develop scalable algorithms was limited to the solution of problems with a given friction.

Though the methodology that was developed for the frictionless contact problems in the previous chapter can be adapted rather easily for 2D problems with Tresca friction [26], this was not the case for the 3D problems due to the circular or elliptic constraints arising in dual formulation. The situation changed with the development of algorithms for the solution of the latter problems with a bound on the rate of convergence in terms of the bounds on the spectrum of the Hessian (see comments in Chaps. 7–9). The optimality results presented in this chapter appeared in Dostál et al. [27]. The latter paper is the result of a graduate development starting with the approximation of circles by the intersections of squares in Haslinger, Kučera, and Dostál [28] and the early results without the optimality theory (see Dostál et al. [24]). Important issues related to the implementation of anisotropic friction are resolved in Bouchala et al. [29].

References

1. Hlaváček, I., Haslinger, J., Nečas, J., Lovíšek, J.: Solution of Variational Inequalities in Mechanics. Springer, Berlin (1988)
2. Kikuchi, N., Oden, J.T.: Contact Problems in Elasticity. SIAM, Philadelphia (1988)
3. Eck, C., Jarůšek, J., Krbec, M.: Unilateral Contact Problems. Chapman & Hall/CRC, London (2005)
4. Haslinger, J., Hlaváček, I., Nečas, J.: Handbook of Numerical Analysis. Numerical Methods for Unilateral Problems in Solid Mechanics, vol. IV, pp. 313–485. North-Holland, Amsterdam (1996)
5. Wohlmuth, B.I.: Variationally consistent discretization scheme and numerical algorithms for contact problems. Acta Numerica, 569–734 (2011)
6. Panagiotopoulos, P.D.: A nonlinear programming approach to the unilateral contact and friction boundary value problem. Ing.-Archiv **44**, 421–432 (1975)
7. Bazaraa, M.S., Shetty, C.M., Sherali, H.D.: Nonlinear Programming. Theory and Algorithms, 2nd edn. Wiley, New York (1993)
8. Kozubek, T., Markopoulos, A., Brzobohatý, T., Kučera, R., Vondrák, V., Dostál, Z.: MatSol–MATLAB efficient solvers for problems in engineering. http://industry.it4i.cz/en/products/matsol/ (2015)
9. Duvaut, G., Lions, J.L.: Inequalities in Mechanics and Physics. Springer, Berlin (1976)
10. Capatina, A.: Variational Inequalities and Frictional Contact Problems. Springer, New York (2014)
11. Laursen, T.: Computational Contact and Impact Mechanics. Springer, Berlin-Heildelberg (2002)
12. Wriggers, P.: Contact Mechanics. Springer, Berlin (2005)
13. Wohlmuth, B.I., Krause, R.: Monotone methods on nonmatching grids for nonlinear contact problems. SIAM J. Sci. Comput. **25**, 324–347 (2003)
14. Krause, R.H.: On the Multiscale Solution of Constrained Minimization Problems. Domain Methods in Science and Engineering XVII. Lecture Notes in Computational Science and Engineering, vol. 60, pp. 93–104. Springer, Berlin (2008)
15. Krause, R.H.: A Non-smooth multiscale method for solving frictional two-body contact problems in 2D and 3D with multigrid efficiency. SIAM J. Sci. Comput. **31**(2), 1399–1423 (2009)
16. Qi, L., Sun, J.: A nonsmooth version of Newton's method. Math. Progr. **58**, 353–367 (1993)
17. Hager, C., Wohlmuth, B.I.: Nonlinear complementarity functions for plasticity problems with frictional contact. Comput. Methods Appl. Mech. Eng. **198**, 3411–3427 (2009)

18. Hintermüller, M., Ito, K., Kunisch, K.: The primal-dual active set strategy as a semi-smooth Newton method. SIAM J. Optim. **13**, 865–888 (2002)
19. Migorski, S., Ochal, A., Sofonea, M.: Modeling and analysis of an antiplane piezoelectric contact problem. Math. Models Methods Appl. Sci. **19**, 1295–1324 (2009)
20. Hüeber, S., Matei, A., Wohlmuth, B.I.: A contact problem for electro-elastic materials. ZAMM - J. Appl. Math. Mech. **93**, 789–800 (2013)
21. Ito, K., Kunisch, K.: On a semi-smooth Newton method and its globalization. Math. Progr. Ser. A **118**, 347–370 (2009)
22. Kučera, R., Motyčková, K., Markopoulos, A.: The R-linear convergence rate of an algorithm arising from the semi-smooth Newton method applied to 2D contact problems with friction. Comput. Optim. Appl. **61**(2), 437–461 (2015)
23. Dostál, Z., Vondrák, V.: Duality based solution of contact problems with Coulomb friction. Arch. Mech. **49**(3), 453–460 (1997)
24. Dostál, Z., Horák, D., Kučera, R., Vondrák, V., Haslinger, J., Dobiáš, J., Pták, S.: FETI based algorithms for contact problems: scalability, large displacements and 3D Coulomb friction. Comput. Methods Appl. Mech. Eng. **194**(2–5), 395–409 (2005)
25. Nečas, J., Jarušek, J., Haslinger, J.: On the solution of the variational inequality to the Signorini problem with small friction. Bollettino dell'Unione Matematica Italiana **5**(17–B), 796–811 (1980)
26. Dostál, Z., Kozubek, T., Markopoulos, A., Brzobohatý, T., Vondrák, V., Horyl, P.: Scalable TFETI algorithm for two dimensional multibody contact problems with friction. J. Comput. Appl. Math. **235**(2), 403–418 (2010)
27. Dostál, Z., Kozubek, T., Markopoulos, A., Brzobohatý, T., Vondrák, V., Horyl, P.: Theoretically supported scalable TFETI algorithm for the solution of multibody 3D contact problems with friction. Comput. Methods Appl. Mech. Eng. 205–208 (2012)
28. Haslinger, J., Kučera, R., Dostál, Z.: An algorithm for the numerical realization of 3D contact problems with Coulomb friction. J. Comput. Appl. Math. **164–165**, 387–408 (2004)
29. Bouchala, J., Dostál, Z., Kozubek, T., Pospíšil, L., Vodstrčil, P.: On the solution of convex QPQC problems with elliptic and other separable constraints. Appl. Math. Comput. **247**(15), 848–864 (2014)

Chapter 13
Transient Contact Problems

In this chapter, we adapt the methods presented above to the solution of contact problems which take inertia forces into account. The introduction of time is a source of many complications, even in the case that we restrict our attention to the 3D problems of linear elasticity without friction. A little is known about a solution of continuous problems, and the cost of the solution increases also with the number of time intervals required by a time discretization. On the other hand, the problem is physically more realistic, as the real world problems are naturally time dependent, and the time discretization reduces the problem to a sequence of quasi-static problems with a good initial approximation. It is even possible to develop an explicit computational scheme that uses the initial approximation at the beginning of each time step to evaluate explicitly an approximation of a solution at the end of the time step [1] at the cost of a large number of time steps. Here, we present an implicit scheme which uses much smaller number of time steps at the cost of solving two well-conditioned bound constrained QP problems.

The FETI-type domain decomposition methods like TFETI comply well with the structure of auxiliary problems arising in the solution of transient contact problems. It turns out that the implementation of the time step problems by TFETI is even simpler than that in Chap. 11 since the effective stiffness matrices of the "floating" subdomains are regularized by the mass matrices and have their condition number uniformly bounded by a constant which is independent of the number of nodal variables associated with the space discretization. To reduce the number of iterations, it is possible to use the preconditioning by a conjugate projector (deflation), which improves the performance of both linear and nonlinear steps of the algorithms.

The basic TFETI-based algorithms presented here are effective for the parallel solution of transient contact problems. There is no costly orthogonalization of dual equality constraints, so that it can use tens of thousands of cores to solve, for a reasonable number of time steps, as large problems as reported in the previous chapters, i.e., tens of thousands of subdomains and billions of nodal variables. For larger problems, the communication can dominate the costs. Some hints for massively parallel implementation can be found in Chap. 19.

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3_13

13.1 Transient Multibody Frictionless Contact Problem

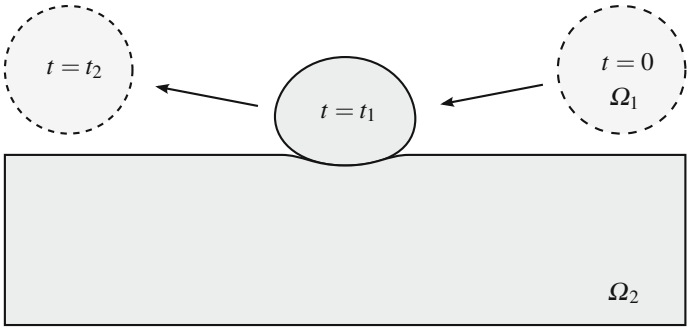

Fig. 13.1 Transient multibody contact problem at the time $t = 0, t_1, t_2$

Let us consider a system of s homogenous isotropic elastic bodies, each of which occupies at time $t = 0$ a reference configuration, a bounded domain $\Omega^p \subset \mathbb{R}^3$, $p = 1, \ldots, s$, with a sufficiently smooth boundary Γ^p. The mechanical properties of Ω^p, which are assumed to be homogeneous, are defined by the Young modulus E^p, the Poisson ratio ν^p, and the density ρ^p. The density defines the inertia forces. On each Ω^p, there is defined the vector of external body forces

$$\mathbf{f}^p : \Omega^p \to \mathbb{R}^d.$$

We suppose that each Γ^p consists of three disjoint parts Γ_U^p, Γ_F^p, and Γ_C^p that do not change in time,

$$\Gamma^p = \overline{\Gamma}_U^p \cup \overline{\Gamma}_F^p \cup \overline{\Gamma}_C^p,$$

and that there are prescribed zero displacements

$$\mathbf{u}^p = \mathbf{o} \quad \text{on} \quad \Gamma_U^p \times [0, T]$$

and traction

$$\mathbf{f}_\Gamma^p : \Gamma_F^p \times [0, T] \to \mathbb{R}^3.$$

We denote

$$\Omega = \cup_{p=1}^s \Omega^p, \quad \Gamma_U = \cup_{p=1}^s \Gamma_U^p, \quad \Gamma_F = \cup_{p=1}^s \Gamma_F^p.$$

The part Γ_C^p denotes the part of Γ^p that can get into contact with other bodies in a time interval $[0, T]$, $T > 0$. We denote by Γ_C^{pq} the part of Γ_C^p that can come in contact with Ω^q in the time interval $[0, T]$. We admit $\Gamma_U^p = \emptyset$ for any $p = 1, \ldots, s$, but we assume that all Γ_C^p are sufficiently smooth so that for almost every $\mathbf{x}^p \in \Gamma_C^p$, there is a unique external normal $\mathbf{n}^p = \mathbf{n}^p(\mathbf{x}^p)$. See also Fig. 13.1. We assume that we can neglect the change of $\mathbf{n}^p$ in time.

The main new features of frictionless transient contact problems, as compared with Sect. 11.2, apart from the time dimension, are the inertia forces and the initial conditions defined by $\mathbf{u}_0 : \Omega \to \mathbb{R}^3$ and $\mathbf{d}_0 : \Omega \to \mathbb{R}^3$.

To describe the non-penetration on Γ_C^{pq}, we choose a contact coupling set $\mathscr{S}$ and for each $(p, q) \in \mathscr{S}$, we define a one-to-one continuous mapping

$$\chi^{pq} : \Gamma_C^{pq} \to \Gamma_C^{qp}$$

as in Sect. 11.2. The *(strong) linearized non-penetration condition* at $\mathbf{x} \in \Gamma_C^{pq}$ and time $t \in [0, T]$ then reads

$$\left(\mathbf{u}^p(\mathbf{x}) - \mathbf{u}^q \circ \chi^{pq}(\mathbf{x})\right) \cdot \mathbf{n}^p(\mathbf{x}) \leq \left(\chi^{pq}(\mathbf{x}) - \mathbf{x}\right) \cdot \mathbf{n}^p(\mathbf{x}). \tag{13.1}$$

Denoting the surface traction on the slave side of the active contact interface without friction by

$$\boldsymbol{\lambda} = -\sigma(\mathbf{u}^p)\mathbf{n}^p = (\boldsymbol{\lambda} \cdot \mathbf{n}^p)\,\mathbf{n}^p, \quad (\mathbf{x}, t) \in \Gamma_C^p \times [0, T],$$

and using the notation of Sect. 11.2, we can write the complete contact conditions for $(p, q) \in \mathscr{S}$ in the form

$$[u_n] \leq g, \quad \lambda_n \geq 0, \quad \lambda_n([u_n] - g) = 0, \quad (\mathbf{x}, t) \in \Gamma_C^{pq} \times [0, T]. \tag{13.2}$$

The equations of motion and the constraints that should be satisfied by the displacements

$$\mathbf{u}^p : \Omega^p \cup \Gamma^p \times [0, T] \to \mathbb{R}^3, \ T > 0,$$

read

$$\rho\ddot{\mathbf{u}} - \operatorname{div}\sigma(\mathbf{u}) = \mathbf{f} \quad \text{in} \quad \Omega \quad \times [0, T], \tag{13.3}$$

$$\mathbf{u} = \mathbf{o} \quad \text{on} \quad \Gamma_U \quad \times [0, T], \tag{13.4}$$

$$\sigma(\mathbf{u})\,\mathbf{n} = \mathbf{f}_\Gamma \quad \text{on} \quad \Gamma_F \quad \times [0, T], \tag{13.5}$$

$$\boldsymbol{\lambda} = \lambda_n \mathbf{n} \quad \text{on} \quad \Gamma_C^{pq} \quad \times [0, T], \quad (p, q) \in \mathscr{S}, \tag{13.6}$$

$$\lambda_n \geq 0 \quad \text{on} \quad \Gamma_C^{pq} \quad \times [0, T], \quad (p, q) \in \mathscr{S}, \tag{13.7}$$

$$[u_n] \leq g \quad \text{on} \quad \Gamma_C^{pq} \quad \times [0, T], \quad (p, q) \in \mathscr{S}, \tag{13.8}$$

$$\lambda_n([u_n] - g) = 0 \quad \text{on} \quad \Gamma_C^{pq} \quad \times [0, T], \quad (p, q) \in \mathscr{S}, \tag{13.9}$$

$$\mathbf{u}(., 0) = \mathbf{u}_0 \quad \text{in} \quad \Omega, \tag{13.10}$$

$$\dot{\mathbf{u}}(., 0) = \mathbf{d}_0 \quad \text{in} \quad \Omega. \tag{13.11}$$

The relations (13.3)–(13.11) represent the general formulation of a transient (dynamic) multibody frictionless contact problem. The relations include Newton's equation of motion (13.3), the classical boundary conditions (13.4)–(13.5), the contact conditions (13.6)–(13.9), and the initial conditions (13.10) and (13.11).

13.2 Variational Formulation and Domain Decomposition

To get a formulation of transient contact problem (13.3)–(13.11) that is suitable for a finite element discretization in space, let us first define the test spaces

$$V^p = \left\{ \mathbf{v}^p \in (H^1\left(\Omega^p\right))^3 : \ \mathbf{v}^p = \mathbf{o} \text{ on } \Gamma_U^p \right\}, \qquad V = V^1 \times \cdots \times V^s.$$

If we multiply the Newton equation of motion (13.3) by $\mathbf{v} \in V$, integrate the result over Ω, and notice that (13.3)–(13.5) differ from the conditions of static equilibrium (11.5) only by the first term in (13.3), we can use procedures of Sect. 11.3 to get

$$\sum_{p=1}^{s} \int_{\Omega^p} \varrho^p \ddot{\mathbf{u}}^p \cdot \mathbf{v}^p \, \mathrm{d}\Omega + a\left(\mathbf{u}, \mathbf{v}\right) + \sum_{(p,q)\in\mathscr{S}} \int_{\Gamma_C^{pq}} \lambda_n [v_n] \, \mathrm{d}\Gamma = \ell(\mathbf{v}), \tag{13.12}$$

where a and ℓ are defined in Sect. 11.3. For a sound variational formulation of the equations of motion (13.3)–(13.5), it is useful to admit $\lambda_n \in M^+$, where

$$M^+ = \prod_{(p,q)\in\mathscr{S}} \left\{ \mu^{pq} \in H^{-\frac{1}{2}}(\Gamma_C^{pq}) : \int_{\Gamma_C^{pq}} \langle \mu^{pq}, [v]_n \rangle \, \mathrm{d}\Gamma \geq 0 \text{ for } \mathbf{v} \in V, \ [v]_n \geq 0 \right\}.$$

In this case, we should replace the second sum in (13.12) by the duality pairing $\langle \lambda_n, [v_n] \rangle$. Using the notation

$$m(\ddot{\mathbf{u}}, \mathbf{v}) = \sum_{p=1}^{s} \int_{\Omega^p} \varrho^p \ddot{\mathbf{u}}^p \cdot \mathbf{v}^p \, \mathrm{d}\Omega, \qquad \ddot{\mathbf{u}}, \mathbf{v} \in V,$$

we can rewrite the variational equations (13.12) in the compact form

$$m\left(\ddot{\mathbf{u}}, \mathbf{v}\right) + a\left(\mathbf{u}, \mathbf{v}\right) + \langle \lambda_n, \mathbf{v} \rangle = \ell(\mathbf{v}). \qquad \mathbf{v} \in V. \tag{13.13}$$

To write the contact conditions (13.6)–(13.9) in a form which admits $\lambda_n \in M^+$, which we shall use in Chap. 15, let us apply $\mu \in M^+$ to the non-penetration condition (13.8) and add the result to (13.9). After a simple manipulation, we get

$$\langle \mu - \lambda_n, [u_n] \rangle \ \leq \ \langle \mu - \lambda_n, g \rangle, \quad \mu \in M^+. \tag{13.14}$$

Thus we got the problem to find for almost every $t \in [0, T]$ a displacement $\mathbf{u}(\cdot\ , t) \in V$ and $\lambda_n \in M^+$ that satisfy (13.13), (13.14), (13.10), and (13.11).

Moreover, we assume that the solution $\mathbf{u}$ is sufficiently smooth so that $\ddot{\mathbf{u}}$ exists in some reasonable sense and can be approximated by finite differences.

A little is known about the solvability of transient problems, so we shall assume in what follows that a solution exists. For a more complete discussion of solvability and variational formulation of dynamic problems, including problems with friction, see Wohlmuth [2] or Eck, Jarůšek, and Krbec [3].

The domain decomposition is similar to that described in Sect. 11.4. We tear each body from the part of the boundary with Dirichlet boundary conditions, decompose each body into subdomains, assign each subdomain a unique number, and introduce new "gluing" conditions on the artificial intersubdomain boundaries and on the boundaries with imposed Dirichlet conditions. Misusing a little the notation, we denote the subdomains and their number again by Ω^p and s, respectively. For the artificial intersubdomain boundaries, we introduce the notation analogously to the notation of the contact boundary, so that Γ_G^{pq} denotes the part of Γ^p that is glued to Ω^q. Obviously $\Gamma_G^{pq} = \Gamma_G^{qp}$. The gluing conditions require the continuity of the displacements and their normal derivatives across the intersubdomain boundaries. An auxiliary decomposition of the problem of Fig. 13.1 with renumbered subdomains and artificial intersubdomain boundaries is similar to that in Fig. 11.3. The procedure is essentially the same as that in Sect. 11.4.

13.3 Discretization

Using the finite element quasi-uniform semi-discretization in space with shape regular elements and the procedure similar to that described in Sect. 11.5, we get a matrix counterpart of (13.13) and (13.6)–(13.11) at the time t

$$\mathsf{M}\ddot{\mathbf{u}} + \mathsf{K}\mathbf{u} = \mathbf{f} - \mathsf{B}_I^T \boldsymbol{\lambda}_I^T - \mathsf{B}_E^T \boldsymbol{\lambda}_E, \tag{13.15}$$

$$\mathsf{B}_I \mathbf{u} \le \mathbf{c}_I, \tag{13.16}$$

$$\mathsf{B}_E \mathbf{u} = \mathbf{c}_E = \mathbf{o}, \tag{13.17}$$

$$\boldsymbol{\lambda}_I \ge \mathbf{o}, \tag{13.18}$$

$$\boldsymbol{\lambda}^T (\mathsf{B}\mathbf{u} - \mathbf{c}) = 0. \tag{13.19}$$

Here, we use the same notation (bold symbol) for the continuous displacements $\mathbf{u} \in V \times [0, T]$ and its vector representation $\mathbf{u} \in \mathbb{R}^n \times [0, T]$. Due to the TFETI domain decomposition, the finite element semi-discretization in space results in the block diagonal stiffness matrix $\mathsf{K} = \operatorname{diag}(\mathsf{K}_1, \ldots, \mathsf{K}_s)$ of the order n with the sparse SPS diagonal blocks K_p that correspond to the subdomains Ω^p. The same structure has the positive definite mass matrix $\mathsf{M} = \operatorname{diag}(\mathsf{M}_1, \ldots, \mathsf{M}_s)$. The decomposition also induces the block structure of the vector $\mathbf{f} \in \mathbb{R}^n \times [0, T]$ of the discretized form ℓ and the vector of nodal displacements $\mathbf{u} \in \mathbb{R}^n \times [0, T]$.

The matrix $\mathsf{B}_I \in \mathbb{R}^{m_I \times n}$ and the vector $\mathbf{c}_I \in \mathbb{R}^{m_I}$ describe the linearized nonpenetration conditions as in Sect. 11.5. Similarly the matrix $\mathsf{B}_E \in \mathbb{R}^{m_E \times n}$ and the vector $\mathbf{c}_E \in \mathbb{R}^{m_E}$ with the entries $c_i = 0$ enforce prescribed zero displacements on

the part of the boundary with imposed Dirichlet conditions and the continuity of displacements across the auxiliary interfaces, respectively. Typically, both m_I and m_E are much smaller than n. We assume that the matrix $\mathsf{B} = [\mathsf{B}_E^T, \mathsf{B}_I^T]^T$ has orthonormal rows so that it satisfies (11.22) (see also Remark 11.1).

Finally, let $\boldsymbol{\lambda}_I \in \mathbb{R}^{m_I}$ and $\boldsymbol{\lambda}_E \in \mathbb{R}^{m_E}$ denote the components of the vector of Lagrange multipliers $\boldsymbol{\lambda} = \boldsymbol{\lambda}_t \in \mathbb{R}^m$, $m = m_I + m_E$, at time t. We use the notation

$$\boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{\lambda}_I \\ \boldsymbol{\lambda}_E \end{bmatrix}, \quad \mathsf{B} = \begin{bmatrix} \mathsf{B}_I \\ \mathsf{B}_E \end{bmatrix}, \quad \text{and} \quad \mathbf{c} = \begin{bmatrix} \mathbf{c}_I \\ \mathbf{c}_E \end{bmatrix}. \tag{13.20}$$

For the time discretization, we use the *contact-stabilized Newmark scheme* introduced by Krause and Walloth [4] with the regular partition of the time interval $[0, T]$,

$$0 = t_0 < t_1 \ldots < t_{n_T} = T, \quad t_k = k\Delta, \quad \Delta = T/n_T, \quad k = 0, \ldots, n_T.$$

The scheme assumes that the acceleration vector is split at time t_k into two components $\ddot{\mathbf{u}}_k^{\text{con}}$ and $\ddot{\mathbf{u}}_k^{\text{int}}$ related to the acceleration affected by the contact and other forces, respectively,

$$\ddot{\mathbf{u}}_k = \ddot{\mathbf{u}}_k^{\text{con}} + \ddot{\mathbf{u}}_k^{\text{int}}, \tag{13.21}$$

where

$$\ddot{\mathbf{u}}_k^{\text{int}} = \mathsf{M}^{-1}(\mathbf{f}_k - \mathsf{K}\mathbf{u}_k) \quad \text{and} \quad \ddot{\mathbf{u}}_k^{\text{con}} = -\mathsf{M}^{-1}\mathsf{B}^T\boldsymbol{\lambda}_k. \tag{13.22}$$

Denoting

$$\mathscr{K} = \{\mathbf{u} \in \mathbb{R}^n : \mathsf{B}_I\mathbf{u} \le \mathbf{c}_I \ \text{ and } \ \mathsf{B}_E\mathbf{u} = \mathbf{o}\},$$

we can write the solution algorithm in the following form.

Algorithm 13.1 Contact-stabilized Newmark algorithm.

Step 0. {*Initialization.*}
Set $\mathbf{u}_0$, $\dot{\mathbf{u}}_0$, $\widetilde{\mathsf{K}} = \frac{4}{\Delta^2}\mathsf{M} + \mathsf{K}$, $\boldsymbol{\lambda}_0^{\text{pred}}$, $T > 0$, $n_T \in \mathbb{N}$, *and* $\Delta = T/n_T$
for $k = 0, \ldots, n_T - 1$
Step 1. {*Predictor displacement.*}
Find the minimizer $\mathbf{u}_{k+1}^{\text{pred}}$ and the multiplier $\boldsymbol{\lambda}_{k+1}^{\text{pred}}$ for
$\min_{\mathbf{u}\in\mathscr{K}} \ \frac{1}{2}\mathbf{u}^T\mathsf{M}\mathbf{u} - \left(\mathsf{M}\mathbf{u}_k + \Delta\mathsf{M}\dot{\mathbf{u}}_k - \mathsf{B}^T\boldsymbol{\lambda}_k^{\text{pred}}\right)^T\mathbf{u}$

Step 2. {*Contact-stabilized displacement.*}
Find the minimizer $\mathbf{u}_{k+1}$ and the multiplier $\boldsymbol{\lambda}_{k+1}$ for
$\min_{\mathbf{u}\in\mathscr{K}} \ \frac{1}{2}\mathbf{u}^T\widetilde{\mathsf{K}}\mathbf{u} - \left(\frac{4}{\Delta^2}\mathsf{M}\mathbf{u}_{k+1}^{\text{pred}} - \mathsf{K}\mathbf{u}_k + \mathbf{f}_k + \mathbf{f}_{k+1} - \mathsf{B}^T\boldsymbol{\lambda}_k\right)^T\mathbf{u}$

Step 3. {*Velocity update.*}
$\dot{\mathbf{u}}_{k+1} = \dot{\mathbf{u}}_k + \frac{2}{\Delta}\left(\mathbf{u}_{k+1} - \mathbf{u}_{k+1}^{\text{pred}}\right)$
end for

The matrix $\widetilde{\mathsf{K}}$ introduced in Step 0 is called the *effective stiffness matrix*.

13.4 Dual Formulation of Time Step Problems

The cost of the iterates of Algorithm 13.1 is dominated by the cost of the execution of its Step 1 and Step 2, the implementation of which requires the solution of bound and equality constrained QP problems

$$\min f(\mathbf{u}) \quad \text{subject to} \quad \mathsf{B}_I\mathbf{u} \le \mathbf{c}_I \quad \text{and} \quad \mathsf{B}_E\mathbf{u} = \mathbf{c}_E \tag{13.23}$$

with the strictly convex quadratic function

$$f(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T\mathsf{H}\mathbf{u} - \mathbf{u}^T\mathbf{h}.$$

The function f for the problem arising in Step 1 is defined by

$$\mathsf{H} = \mathsf{M} \quad \text{and} \quad \mathbf{h} = \mathsf{M}\mathbf{u}_k + \Delta\mathsf{M}\dot{\mathbf{u}}_k - \mathsf{B}^T\boldsymbol{\lambda}_k^{\text{pred}}$$

and for that arising in Step 2 by

$$\mathsf{H} = \widetilde{\mathsf{K}} \quad \text{and} \quad \mathbf{h} = \frac{4}{\Delta^2}\mathsf{M}\mathbf{u}_{k+1}^{\text{pred}} - \mathsf{K}\mathbf{u}_k + \mathbf{f}_k + \mathbf{f}_{k+1} - \mathsf{B}^T\boldsymbol{\lambda}_k.$$

Even though (13.23) is a standard QP problem, its form is not suitable for numerical solution due to general inequality constraints. This complication can be essentially reduced by applying the duality theory of convex programming (see Sect. 3.7) as in Sect. 11.6. In the dual formulation of problem (13.23), we use the Lagrange multipliers with two block components, namely $\boldsymbol{\lambda}_I \in \mathbb{R}^{m_I}$ associated with the nonpenetration condition and $\boldsymbol{\lambda}_E \in \mathbb{R}^{m_E}$ associated with the "gluing" and prescribed displacements, so the Lagrangian associated with problem (13.23) reads

$$L(\mathbf{u}, \boldsymbol{\lambda}) = f(\mathbf{u}) + \boldsymbol{\lambda}^T(\mathsf{B}\mathbf{u} - \mathbf{c}). \tag{13.24}$$

Using Proposition 3.13 (with $\mathsf{R} = [\mathbf{o}]$), we get that $\boldsymbol{\lambda}$ solves the strictly convex minimization problem

$$\min \ \frac{1}{2}\boldsymbol{\lambda}^T\mathsf{B}\mathsf{H}^{-1}\mathsf{B}^T\boldsymbol{\lambda} - \boldsymbol{\lambda}^T(\mathsf{B}\mathsf{H}^{-1}\mathbf{h} - \mathbf{c}) \quad \text{s.t.} \quad \boldsymbol{\lambda}_I \ge \mathbf{o}. \tag{13.25}$$

Once the solution $\widehat{\boldsymbol{\lambda}}$ of (13.25) is known, the solution $\widehat{\mathbf{u}}$ of (13.23) may be evaluated from the KKT condition

$$\mathsf{H}\mathbf{u} - \mathbf{h} + \mathsf{B}^T\widehat{\boldsymbol{\lambda}} = \mathbf{o}.$$

Problem (13.25) is much more suitable for computations than (13.23) because we replaced the general inequality constraints in (13.23) by nonnegativity constraints. The favorable distribution of the spectrum of the mass matrix M and the nonnegativity

constraints are even sufficient to implement Step 1 by the standard MPRGP algorithm with the asymptotically linear complexity. The estimates in the next section show that similar relations hold also for Step 2.

13.5 Bounds on the Spectrum of Dual Energy Function

To simplify the notation, let us denote

$$\widetilde{\mathsf{F}} = \mathsf{B}\widetilde{\mathsf{K}}^{-1}\mathsf{B}^T \text{ and } \mathbf{d} = \mathsf{B}\widetilde{\mathsf{K}}^{-1}\mathbf{h} - \mathbf{c},$$

so problem (13.25) with $\mathsf{H} = \widetilde{\mathsf{F}}$ reads

$$\min \ \Theta(\boldsymbol{\lambda}) \quad \text{s.t. } \boldsymbol{\lambda} \in \Omega_B, \tag{13.26}$$

where

$$\Theta(\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\lambda}^T\widetilde{\mathsf{F}}\boldsymbol{\lambda} - \boldsymbol{\lambda}^T\mathbf{d} \quad \text{and} \quad \Omega_B = \{\boldsymbol{\lambda} \in \mathbb{R}^m : \ \boldsymbol{\lambda}_I \geq \mathbf{o}\}. \tag{13.27}$$

In order to avoid a proliferation of constants, we shall use the notation $A \lesssim B$ (or $B \lesssim A$) introduced by Brenner [5] to represent the statement that there are constants C_1 and C_2 independent of h, H, Δ, and other variables that appear on both sides such that $A \leq C_1 B$ (or $B \leq C_2 A$). The notation $A \approx B$ means that $A \lesssim B$ and $B \lesssim A$. Thus the quadratic form defined by the mass matrix satisfies

$$\mathbf{x}^T\mathsf{M}\mathbf{x} \approx h^3\|\mathbf{x}\|^2, \quad \mathbf{x} \in \mathbb{R}^n \tag{13.28}$$

(see, e.g., Wathen [6]). We also need a variant of the standard estimate [7].

Lemma 13.1 *Let* K *denote the finite element stiffness matrix of Sect. 13.3 arising from a quasi-uniform discretization of the subdomains* Ω^p, $p = 1, \ldots, s$, *using linear shape regular tetrahedral elements* ω_i^p, $i = 1, \ldots, n_p$ *with the discretization parameter* h. *Then*

$$\|\mathsf{K}\| \lesssim h. \tag{13.29}$$

Proof First observe that we assume that the components c^p_{ijkl} of Hooke's tensor of elasticity C^p are bounded, so the energy density produced by the strain $\boldsymbol{\varepsilon}(\mathbf{v}^p)$ associated with the displacement $\mathbf{v}^p$ in the subdomain Ω^p satisfies

$$\sigma(\mathbf{v}^p) : \boldsymbol{\varepsilon}(\mathbf{v}^p) \lesssim \|\boldsymbol{\varepsilon}(\mathbf{v}^p)\|_F^2 = \sum_{i,j=1}^{3} e_{ij}^2(\mathbf{v}^p).$$

Let $\boldsymbol{\phi}_1^p, \ldots, \boldsymbol{\phi}_{n_p}^p$ denote a finite element basis of the subspace V_h^p of $\left(H^1(\Omega^p)\right)^3$ that satisfies the assumptions, $p = 1, \ldots, s$, so that any $\mathbf{u}_h^p$ can be written in the form

$$\mathbf{u}_h^p = \xi_1 \boldsymbol{\phi}_1^p + \cdots + \xi_{n_p} \boldsymbol{\phi}_{n_p}^p,$$

and observe that

$$\|\boldsymbol{\varepsilon}(\boldsymbol{\phi}_i^p)\|_F^2 \lesssim h^{-2}.$$

Thus the elements k_{ij} of K_p satisfy

$$k_{ij}^p = \int_{\omega_i^p \cap \omega_j^p} a(\boldsymbol{\phi}_i^p, \boldsymbol{\phi}_j^p)\, \mathrm{d}\omega = \int_{\omega_i^p \cap \omega_j^p} \sigma(\boldsymbol{\phi}_i^p) : \boldsymbol{\varepsilon}(\boldsymbol{\phi}_j^p)\, \mathrm{d}\omega \lesssim h^3 h^{-2} = h.$$

Taking into account that $k_{ij}^p = 0$ for $\omega_i^p \cap \omega_j^p = \emptyset$, we get

$$\boldsymbol{\xi}^T \mathsf{K}_p \boldsymbol{\xi} \lesssim h \|\boldsymbol{\xi}^T\|^2.$$

To complete the proof, just notice that $\mathsf{K} = \operatorname{diag}(\mathsf{K}_1, \ldots, \mathsf{K}_s)$. □

Lemma 13.2 *Let the assumptions of Lemma 13.1 hold, $C > 0$, and*

$$\|\mathsf{B}^T \boldsymbol{\lambda}\|^2 \approx \|\boldsymbol{\lambda}\|^2.$$

Then

$$\frac{h^2 \Delta^2}{h^3 \left(h^2 + \Delta^2\right)} \|\boldsymbol{\lambda}\|^2 \lesssim \boldsymbol{\lambda}^T \widetilde{\mathsf{F}} \boldsymbol{\lambda} \lesssim \frac{\Delta^2}{h^3} \|\boldsymbol{\lambda}\|^2. \tag{13.30}$$

Proof Let $\boldsymbol{\lambda} \in \mathbb{R}^m$, let $\mu_{\min}$ and $\mu_{\max}$ denote the extreme eigenvalues of M, and let $\boldsymbol{\mu} = \mathsf{B}^T \boldsymbol{\lambda}$. Then we have $\mu_{\min} \approx h^3$ by (13.28). Since the smallest eigenvalue of K is zero, we have

$$\begin{aligned} \boldsymbol{\lambda}^T \widetilde{\mathsf{F}} \boldsymbol{\lambda} = \boldsymbol{\mu}^T \widetilde{\mathsf{K}}^{-1} \boldsymbol{\mu} &= \boldsymbol{\mu}^T \left(\mathsf{K} + \frac{4}{\Delta^2} \mathsf{M}\right)^{-1} \boldsymbol{\mu} \\ &\leq \frac{\Delta^2}{4\mu_{\min}} \|\boldsymbol{\mu}\|^2 \approx \frac{\Delta^2}{h^3} \|\mathsf{B}^T \boldsymbol{\lambda}\|^2 \approx \frac{\Delta^2}{h^3} \|\boldsymbol{\lambda}\|^2. \end{aligned}$$

Similarly, using Lemma 13.1 and the assumptions, we get

$$\begin{aligned} \boldsymbol{\lambda}^T \widetilde{\mathsf{F}} \boldsymbol{\lambda} = \boldsymbol{\mu}^T \widetilde{\mathsf{K}}^{-1} \boldsymbol{\mu} &= \boldsymbol{\mu}^T \left(\mathsf{K} + \frac{4}{\Delta^2} \mathsf{M}\right)^{-1} \boldsymbol{\mu} \\ &\gtrsim \left(h + h^3 \Delta^{-2}\right)^{-1} \|\boldsymbol{\mu}\|^2 \gtrsim h^{-3} \left(h^{-2} + \Delta^{-2}\right)^{-1} \|\boldsymbol{\lambda}\|^2. \end{aligned}$$

After simple manipulations, we get (13.30). □

The main result of this section is now an easy corollary of Lemma 13.2.

Proposition 13.1 *Let the assumptions of Lemma 13.1 hold, $C > 0$,*

$$\|\mathsf{B}^T\boldsymbol{\lambda}\|^2 \approx \|\boldsymbol{\lambda}\|^2.$$

Then the condition number $\kappa(\widetilde{\mathsf{F}})$ of $\widetilde{\mathsf{F}}$ generated with the discretization parameters h and Δ such that $0 < \Delta \le Ch$ satisfies

$$\kappa(\widetilde{\mathsf{F}}) \lesssim 1. \tag{13.31}$$

13.6 Preconditioning by Conjugate Projector

Examining the proof of Lemma 13.2, we can see that if we use longer time steps, the effective stiffness matrix has very small eigenvalues which obviously correspond to the eigenvectors that are near the kernel of K. This observation was first exploited for linear problems by Farhat, Chen, and Mandel [8] who used the conjugate projectors to the natural coarse grid to achieve scalability with respect to the time step. Unfortunately, this idea cannot be applied to the full extent to the contact problems as we do not know a priori which boundary conditions are applied to the subdomains with nonempty contact interfaces. However, we can still define a preconditioning by the trace of rigid body motions on artificial subdomain interfaces. To implement this observation, we use the preconditioning by conjugate projector for partially constrained strictly convex quadratic programming problems that complies with the MPRGP algorithm for the solution of strictly convex bound constrained problems. Our approach is based on Domorádová and Dostál [9]. Even though the method presented here is similar to that used in Chap. 11, it is not identical, as here we work with the effective stiffness matrix which is nonsingular.

The choice of projectors is motivated by the classical FETI results by Farhat, Mandel, and Roux [10] and their application to the unconstrained transient problems by Farhat, Chen, and Mandel [8]. To explain the motivation in more detail, let us denote by $\mathsf{R} \in \mathbb{R}^{n\times r}$ the full rank matrix, the columns of which whose columns span the kernel of the stiffness matrix K introduced in Sect. 13.2, i.e.,

$$\mathsf{K}\mathsf{R} = \mathsf{O} \quad \text{and} \quad \widetilde{\mathsf{K}}\mathsf{R} = \frac{4}{\Delta^2}\mathsf{M}\mathsf{R}.$$

Since all the subdomains are floating the matrix R is known a priori.

In the static case, we get the same minimization problem as in (13.23) but with $\mathsf{H} = \mathsf{K}$ and $\mathbf{h} = \mathbf{f}$. In this case, H is SPS, so by Proposition 3.13 the dual problem reads

$$\min \quad \frac{1}{2}\boldsymbol{\lambda}^T\mathsf{F}\boldsymbol{\lambda} - \boldsymbol{\lambda}^T\widetilde{\mathbf{d}} \quad \text{s.t.} \quad \boldsymbol{\lambda}_I \ge \mathbf{o} \quad \text{and} \quad \mathsf{G}\boldsymbol{\lambda} = \mathbf{e}, \tag{13.32}$$

where

$$\mathsf{F} = \mathsf{B}\mathsf{K}^{+}\mathsf{B}^{T}, \qquad \widetilde{\mathbf{d}} = \mathsf{B}\mathsf{K}^{+}\mathbf{f} - \mathbf{c},$$
$$\mathsf{G} = \mathsf{T}\mathsf{R}^{T}\mathsf{B}^{T}, \qquad \mathbf{e} = \mathsf{T}\mathsf{R}^{T}\mathbf{f},$$

and T and K^{+} denote a regular matrix that defines the orthonormalization of the rows of $\mathsf{R}^{T}\mathsf{B}^{T}$ and arbitrary generalized inverse to K satisfying $\mathsf{K} = \mathsf{K}\mathsf{K}^{+}\mathsf{K}$, respectively.

The final step in the static case is based on the observation that problem (13.32) is equivalent to

$$\min \quad \frac{1}{2}\boldsymbol{\lambda}^{T}(\mathsf{P}\mathsf{F}\mathsf{P} + \rho\mathsf{Q})\boldsymbol{\lambda} - \boldsymbol{\lambda}^{T}\mathsf{P}\mathbf{d} \quad \text{s.t.} \quad \mathsf{G}\boldsymbol{\lambda} = \mathbf{o} \quad \text{and} \quad \boldsymbol{\lambda}_{I} \geq \ell_{I}, \tag{13.33}$$

where ρ is an arbitrary positive constant,

$$\mathsf{Q} = \mathsf{G}^{T}\mathsf{G} \qquad \text{and} \qquad \mathsf{P} = \mathsf{I} - \mathsf{Q}$$

denote the orthogonal projectors onto the image space of G^{T} and onto the kernel of G, respectively, $\mathbf{d} = \widetilde{\mathbf{d}} - \mathsf{F}\widetilde{\boldsymbol{\lambda}}$, and $\ell = -\widetilde{\boldsymbol{\lambda}}$, where $\widetilde{\boldsymbol{\lambda}}$ is a particular solution of the equation $\mathsf{G}\boldsymbol{\lambda} = \mathbf{e}$ used for the homogenization as in Sect. 11.6.

The proof of optimality for static problems in Sect. 11.10 was based on a favorable distribution of the spectrum of $\mathsf{F}|\mathrm{Ker}\mathsf{G}$. Since $\widetilde{\mathsf{F}}$ can be considered for a large Δ as a perturbation of F, it seems natural to assume that the reduction of iterations to $\mathrm{Ker}\mathsf{G}$ guarantees also some preconditioning effect for $\widetilde{\mathsf{F}}$.

To develop an efficient algorithm for (13.25) based on MPRGP, we have to find a subspace of $\mathbb{R}^{m}$ that is near to $\mathrm{Ker}\mathsf{G}$ and is invariant with respect to the Euclidean projection onto the feasible set. The latter condition implements the requirement for the convergence of multigrid formulated by Iontcheva and Vassilevski [11] that the coarse grid should be defined by the subspace which is kept away from the contact interface. A natural choice $U = \mathrm{Im}\mathsf{U}^{T}$ is defined by the full rank matrix $\mathsf{U} \in \mathbb{R}^{r\times m}$,

$$\mathsf{U} = [\mathsf{O}, \mathsf{G}_{E}], \quad \mathsf{G}_{E} = \mathsf{R}^{T}\mathsf{B}_{E}^{T}.$$

We implement this idea by means of the conjugate projectors

$$\mathsf{P} = \mathsf{I} - \mathsf{U}^{T}(\mathsf{U}\widetilde{\mathsf{F}}\mathsf{U}^{T})^{-1}\mathsf{U}\widetilde{\mathsf{F}} \quad \text{and} \quad \mathsf{Q} = \mathsf{U}^{T}(\mathsf{U}\widetilde{\mathsf{F}}\mathsf{U}^{T})^{-1}\mathsf{U}\widetilde{\mathsf{F}} \tag{13.34}$$

onto the subspaces

$$V = \mathrm{Im}\mathsf{P} \quad \text{and} \quad U = \mathrm{Im}\mathsf{Q} = \mathrm{Im}\mathsf{U}^{T},$$

respectively. It is easy to check directly that P and Q are conjugate projectors, i.e.,

$$\mathsf{P}^{2} = \mathsf{P}, \quad \mathsf{Q}^{2} = \mathsf{Q}, \quad \text{and} \quad \mathsf{P}^{T}\widetilde{\mathsf{F}}\mathsf{Q} = \mathsf{O},$$

and that

$$\mathsf{P}^T\widetilde{\mathsf{F}} = \mathsf{P}^T\widetilde{\mathsf{F}}\mathsf{P} = \widetilde{\mathsf{F}}\mathsf{P} \quad \text{and} \quad \mathsf{Q}^T\widetilde{\mathsf{F}} = \mathsf{Q}^T\widetilde{\mathsf{F}}\mathsf{Q} = \widetilde{\mathsf{F}}\mathsf{Q}. \tag{13.35}$$

From the definition of V and (13.35), we get immediately

$$\mathsf{P}^T\widetilde{\mathsf{F}}\mathsf{P}(\widetilde{\mathsf{F}}V) \subseteq \widetilde{\mathsf{F}}V. \tag{13.36}$$

Thus $\widetilde{\mathsf{F}}V$ is an invariant subspace of $\mathsf{P}^T\widetilde{\mathsf{F}}\mathsf{P}$. The following simple lemma shows that the mapping which assigns each $\boldsymbol{\lambda} \in \widetilde{\mathsf{F}}V$ the vector $\mathsf{P}\boldsymbol{\lambda} \in V$ is expansive.

Lemma 13.3 *Let* P *denote a conjugate projector onto* V. *Then for any* $\boldsymbol{\lambda} \in \widetilde{\mathsf{F}}V$

$$\|\mathsf{P}\boldsymbol{\lambda}\| \geq \|\boldsymbol{\lambda}\| \tag{13.37}$$

and

$$V = \mathsf{P}(\widetilde{\mathsf{F}}V). \tag{13.38}$$

Proof See [12]. □

Using the projector Q, it is possible to solve the auxiliary problem

$$\min_{\boldsymbol{\xi}\in U} \Theta(\boldsymbol{\xi}) = \min_{\boldsymbol{\mu}\in\mathbb{R}^r} \Theta(\mathsf{U}^T\boldsymbol{\mu}) = \min_{\boldsymbol{\mu}\in\mathbb{R}^r} \frac{1}{2}\boldsymbol{\mu}^T\mathsf{U}\widetilde{\mathsf{F}}\mathsf{U}^T\boldsymbol{\mu} - \mathbf{d}^T\mathsf{U}^T\boldsymbol{\mu},$$

with Θ defined in (13.27). By the gradient argument, we get that the minimizer $\widehat{\boldsymbol{\xi}} = \mathsf{U}^T\widehat{\boldsymbol{\mu}}$ of Θ over U is defined by

$$\mathsf{U}\widetilde{\mathsf{F}}\mathsf{U}^T\boldsymbol{\mu} = \mathsf{U}\mathbf{d}, \tag{13.39}$$

so that

$$\widehat{\boldsymbol{\xi}} = \mathsf{U}^T(\mathsf{U}\widetilde{\mathsf{F}}\mathsf{U}^T)^{-1}\mathsf{U}\mathbf{d} = \mathsf{Q}\widetilde{\mathsf{F}}^{-1}\mathbf{d}. \tag{13.40}$$

Thus we can find the minimum of Θ over U effectively whenever we are able to solve (13.39).

We shall use the conjugate projectors Q and P to decompose the minimization problem (13.25) into the minimization on U and the minimization on $V \cap \Omega_B$, $V = \mathrm{Im}\mathsf{P}$. In particular, we shall use three observations. First, using Lemma 13.3, we get that the mapping which assigns to each $\mathbf{x} \in \widetilde{\mathsf{F}}V$ a vector $\mathsf{P}\mathbf{x} \in V$ is an isomorphism. Second, using the definitions of the projectors P and Q, (13.35), and (13.40), we get

$$\mathbf{g}^0 = \widetilde{\mathsf{F}}\widehat{\boldsymbol{\xi}} - \mathbf{d} = \widetilde{\mathsf{F}}\mathsf{Q}\widetilde{\mathsf{F}}^{-1}\mathbf{d} - \mathbf{d} = \mathsf{Q}\widetilde{\mathsf{F}}\widetilde{\mathsf{F}}^{-1}\mathbf{d} - \mathbf{d} = \mathsf{Q}^T\mathbf{d} - \mathbf{d} = -\mathsf{P}^T\mathbf{d}. \tag{13.41}$$

Since

$$\mathrm{Im}\mathsf{P}^T = \mathrm{Im}(\mathsf{P}^T\widetilde{\mathsf{F}}) = \mathrm{Im}(\widetilde{\mathsf{F}}\mathsf{P}) = \widetilde{\mathsf{F}}V \tag{13.42}$$

and $\mathbf{g}^0 \in \mathrm{Im}\mathsf{P}^T$ by (13.41), we get that $\mathbf{g}^0 \in \widetilde{\mathsf{F}}V$. Finally, observe that it follows by Lemma 13.3 that the restriction $\mathsf{P}^T\widetilde{\mathsf{F}}\mathsf{P}|\widetilde{\mathsf{F}}V$ is positive definite.

For any vector $\boldsymbol{\lambda} \in \mathbb{R}^m$, let $\mathscr{I} \subseteq \{1, \ldots, m\}$ denote the set of indices that correspond to $\boldsymbol{\lambda}_I$, so that $\boldsymbol{\lambda}_{\mathscr{I}} = \boldsymbol{\lambda}_I$. Due to our special choice of U, we get that for any $\boldsymbol{\lambda} \in \mathbb{R}^m$

$$[\mathsf{P}\boldsymbol{\lambda}]_{\mathscr{I}} = \boldsymbol{\lambda}_{\mathscr{I}}. \tag{13.43}$$

Moreover, for any $\boldsymbol{\xi} \in U$ and $\boldsymbol{\mu} \in V$, $[\boldsymbol{\xi} + l.\mu]_{\mathscr{I}} \geq \mathbf{o}$ if and only if $\boldsymbol{\mu}_{\mathscr{I}} \geq \mathbf{o}$. Using (13.40), (13.41), and (13.43), we get

$$\begin{aligned}
\min_{\boldsymbol{\lambda}\in\Omega_B} \Theta(\boldsymbol{\lambda}) &= \min_{\substack{\boldsymbol{\xi}\in U,\boldsymbol{\mu}\in V\\ \boldsymbol{\xi}+\boldsymbol{\mu}\in\Omega_B}} \Theta(\boldsymbol{\xi}+\boldsymbol{\mu}) = \min_{\boldsymbol{\xi}\in U} \Theta(\boldsymbol{\xi}) + \min_{\boldsymbol{\mu}\in V\cap\Omega_B} \Theta(\boldsymbol{\mu}) \\
&= \Theta(\widehat{\boldsymbol{\xi}}) + \min_{\boldsymbol{\mu}\in V\cap\Omega_B} \Theta(\boldsymbol{\mu}) = \Theta(\widehat{\boldsymbol{\xi}}) + \min_{\substack{\boldsymbol{\mu}\in\widetilde{\mathsf{F}}V\\ \boldsymbol{\mu}_{\mathscr{I}}\geq\mathbf{o}}} \frac{1}{2}\boldsymbol{\mu}^T\mathsf{P}^T\widetilde{\mathsf{F}}\mathsf{P}\boldsymbol{\mu} - \mathbf{d}^T\mathsf{P}\boldsymbol{\mu} \\
&= \Theta(\widehat{\boldsymbol{\xi}}) + \min_{\substack{\boldsymbol{\mu}\in\widetilde{\mathsf{F}}V\\ \boldsymbol{\mu}_{\mathscr{I}}\geq\mathbf{o}}} \frac{1}{2}\boldsymbol{\mu}^T\mathsf{P}^T\widetilde{\mathsf{F}}\mathsf{P}\boldsymbol{\mu} + \left(\mathbf{g}^0\right)^T\boldsymbol{\mu}.
\end{aligned}$$

Thus we have reduced our bound constrained problem (13.25), after replacing $\boldsymbol{\mu}$ by $\boldsymbol{\lambda}$, to the problem

$$\min_{\substack{\boldsymbol{\lambda}\in\widetilde{\mathsf{F}}V\\ \boldsymbol{\lambda}_I\geq\mathbf{o}}} \frac{1}{2}\boldsymbol{\lambda}^T\mathsf{P}^T\widetilde{\mathsf{F}}\mathsf{P}\boldsymbol{\lambda} + \left(\mathbf{g}^0\right)^T\boldsymbol{\lambda}. \tag{13.44}$$

The following lemma shows that the above problem can be solved by the standard MPRGP algorithm without any change.

Lemma 13.4 *Let $\boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2, \ldots$ be generated by the MPRGP algorithm for the problem*

$$\min_{\boldsymbol{\lambda}_{\mathscr{I}}\geq\mathbf{o}} \frac{1}{2}\boldsymbol{\lambda}^T\mathsf{P}^T\widetilde{\mathsf{F}}\mathsf{P}\boldsymbol{\lambda} + \left(\mathbf{g}^0\right)^T\boldsymbol{\lambda} \tag{13.45}$$

starting from $\boldsymbol{\lambda}^0 = P_{\Omega_B}\left(\mathbf{g}^0\right)$, where P_{Ω_B} denotes the Euclidean projection to the feasible set Ω_B. Then $\boldsymbol{\lambda}^k \in \widetilde{\mathsf{F}}V$, $k = 0, 1, 2, \ldots$.

Proof See Domorádová and Dostál [9] or the book [13]. □

To discuss at least briefly the preconditioning effect of the restriction of the iterates to $\widetilde{\mathsf{F}}V$, let $\varphi_{\min}$ and $\varphi_{\max}$ denote the extreme eigenvalues of $\widetilde{\mathsf{F}}$. Let $\boldsymbol{\mu} \in \widetilde{\mathsf{F}}V$ and $\|\boldsymbol{\mu}\| = 1$. Then by Lemma 13.3

$$\boldsymbol{\mu}^T\mathsf{P}^T\widetilde{\widetilde{\mathsf{F}}}\mathsf{P}\boldsymbol{\mu} = (\mathsf{P}\boldsymbol{\mu})^T\widetilde{\mathsf{F}}(\mathsf{P}\boldsymbol{\mu}) \geq (\mathsf{P}\boldsymbol{\mu})^T\widetilde{\mathsf{F}}(\mathsf{P}\boldsymbol{\mu})/\|\mathsf{P}\boldsymbol{\mu}\|^2 \geq \varphi_{\min} \tag{13.46}$$

and

$$\boldsymbol{\mu}^T \mathsf{P}^T \widetilde{\mathsf{F}} \mathsf{P} \boldsymbol{\mu} \le \boldsymbol{\mu}^T \mathsf{Q}^T \widetilde{\mathsf{F}} \mathsf{Q} \boldsymbol{\mu} + \boldsymbol{\mu}^T \mathsf{P}^T \widetilde{\mathsf{F}} \mathsf{P} \boldsymbol{\mu} = \boldsymbol{\mu}^T \widetilde{\mathsf{F}} \boldsymbol{\mu} \le \varphi_{\max}. \tag{13.47}$$

Thus, it is natural to assume that the preconditioning by the conjugate projector results in an improved rate of convergence. The preconditioning effect can be evaluated by means of the gap between the subspace defined by the eigenvectors corresponding to the smallest eigenvalues of $\widetilde{\mathsf{F}}$ and the subspace that is defined by the image space of U^T (see Dostál [12] or [13]). Let us stress that the unique feature of the preconditioning by a conjugate projector is that it preconditions not only linear steps but also the nonlinear gradient projection steps.

Let us compare our analysis with Farhat, Chen, and Mandel [8], who show that in the unconstrained case, the conjugate projector for large time steps approaches the orthogonal projector to the natural coarse grid (defined by the kernel of G). This observation implies that the convergence properties of their algorithm for longer time steps are similar to the convergence of FETI1. Unfortunately, these arguments are not valid for our algorithm. The reason is that our coarse space is smaller. Indeed, if $\boldsymbol{\lambda} = [\boldsymbol{\lambda}_I^T, \mathbf{o}^T]^T$, then obviously $\boldsymbol{\lambda} \in \mathrm{Ker}\mathsf{U}$, but not necessarily $\boldsymbol{\lambda} \in \mathrm{Ker}\mathsf{G}$. On the other hand, the size of $\boldsymbol{\lambda}_I$ is typically rather small as compared with the size of $\boldsymbol{\lambda}$, so that it is natural to assume that the preconditioning by the conjugate projector improves significantly the rate of convergence.

13.7 Optimality

Now we are ready to show that MPRGP can be used to implement the time step of the implicit Newmark method with a uniformly bounded number of matrix–vector multiplications. Let us consider a class of problems arising from the finite element discretization of a given transient contact problem (13.13) and (13.6)–(13.11) with varying time steps $\Delta > 0$ and the decomposition and discretization parameters H and h, respectively. We assume that the space discretization satisfies the assumptions of Theorem 11.1. Given the constants $C_1, C_2 > 0$, we shall define

$$\mathscr{T}_{C_1 C_2} = \{(H, h, \Delta) \in \mathbb{R}^3 : \ 0 < \Delta \le C_1 h \text{ and } \ h \le C_2\}$$

as the set of indices.

For any $t \in \mathscr{T}_{C_1, C_2}$, we shall define

$$\widetilde{\mathsf{F}}_t = \widetilde{\mathsf{F}}, \quad \mathbf{d}_t = \mathbf{d}, \quad \mathscr{I}_t = \mathscr{I},$$

by the entities generated with the parameters H, h, and Δ, so that problem (13.25) is equivalent to the problem

$$\min \ \frac{1}{2} \boldsymbol{\lambda}_t^T \widetilde{\mathsf{F}}_t \boldsymbol{\lambda}_t - \mathbf{d}_t^T \boldsymbol{\lambda}_t \ \text{ s.t. } \ [\boldsymbol{\lambda}_t]_{\mathscr{I}_t} \ge \mathbf{o}. \tag{13.48}$$

The main result reads as follows.

Theorem 13.1 *Let $C_1, C_2, \varepsilon > 0$, and $0 < c < 1$ denote given constants. For each $t \in \mathcal{T}_{C_1C_2}$, let $\{\boldsymbol{\lambda}_t^k\}$ be generated by Algorithm 8.2 (MPRGP) defined for the solution of problem (13.48) with*

$$\Gamma > 0 \quad \text{and} \quad c\|\widetilde{\mathsf{F}}_t\|^{-1} \le \alpha_t < (2-c)\|\widetilde{\mathsf{F}}_t\|^{-1}.$$

Then MPRGP finds an approximate solution $\widetilde{\boldsymbol{\lambda}}_t$ which satisfies

$$\|\mathbf{g}^P(\widehat{\boldsymbol{\lambda}}_t)\| \le \varepsilon\|\mathbf{d}_t\| \tag{13.49}$$

at $O(1)$ matrix–vector multiplications by the Hessian $\widetilde{\mathsf{F}}_t$ of the cost function.

Proof We shall show that the contraction coefficient of MPGP

$$\eta_{\Gamma,t} = 1 - \frac{\widehat{\alpha}_t\lambda_{\min}(\widetilde{\mathsf{F}}_t)}{\vartheta_t + \vartheta_t\widehat{\Gamma}^2}, \qquad \widehat{\Gamma} = \max\{\Gamma, \Gamma^{-1}\},$$
$$\vartheta_t = 2\max\{\alpha_t\|\widetilde{\mathsf{F}}_t\|, 1\}, \quad \widehat{\alpha}_t = \min\{\alpha_t, 2\|\widetilde{\mathsf{F}}_t\|^{-1} - \alpha_t\},$$

which was established in Theorem 8.1, is bounded away from 1. First recall that by Proposition 13.1 there is a constant $C \ge 1$ such that

$$\kappa(\widetilde{\mathsf{F}}_t) \le C, \quad t \in \mathcal{T}_{C_1C_2}.$$

Since

$$\widehat{\alpha}_t = \min\{\alpha_t, 2\|\widetilde{\mathsf{F}}_t\|^{-1} - \alpha_t\} \ge \min\{c\|\widetilde{\mathsf{F}}_t\|^{-1}, (2-c)\|\widetilde{\mathsf{F}}_t\|^{-1}\} \tag{13.50}$$
$$= c\|\widetilde{\mathsf{F}}_t\|^{-1} > 0, \tag{13.51}$$

it follows that

$$\widehat{\alpha}_t\lambda_{\min}(\widetilde{\mathsf{F}}_t) \ge c\|\widetilde{\mathsf{F}}_t\|^{-1}\lambda_{\min}(\widetilde{\mathsf{F}}_t) = c\kappa(\widetilde{\mathsf{F}}_t)^{-1} \ge c/C.$$

Observing that $\theta_t \le 4$, we get

$$\eta_{\Gamma,t} = 1 - \frac{\widehat{\alpha}_t\lambda_{\min}(\widetilde{\mathsf{F}}_t)}{\vartheta_t + \vartheta_t\widehat{\Gamma}^2} \le 1 - \frac{c}{4C(1+\widehat{\Gamma}^2)} < 1.$$

Thus we have found a nontrivial bound on $\eta_{\Gamma,t}$ which does not depend on H, h, Δ. The rest follows Theorem 8.1. □

Let us finish by recalling that to achieve any kind of scalability, it is important to keep H/h bounded in order to keep asymptotically linear complexity of the matrix decomposition of $\widetilde{\mathsf{K}}$ that is necessary for the effective evaluation of the action of $\widetilde{\mathsf{F}}$.

Notice that the optimality concerns also the norm of the projected gradients. The rate of convergence can be improved by the preconditioning by the conjugate projector.

13.8 Numerical Experiments

The algorithms presented here were implemented in several software packages (see Sect. 19.5) and tested on a number of academic benchmarks and real world problems. Here, we give some results that illustrate their numerical scalability and effectiveness using `MatSol` [14], postponing the demonstration of parallel scalability to Chap. 19. All computations were carried out with the parameters recommended in the description of the algorithms in Chap. 8, i.e., $\Gamma = 1$ and $\alpha = 1.9|\widetilde{\mathsf{F}}|^{-1}$. The relative precision of the computations was 10^{-4}.

13.8.1 Academic Benchmark

We first tested the performance of the TFETI-based algorithms on 3D impact of the elastic bodies Ω_1 and Ω_2 depicted in Fig. 13.2 (see [15]). The top view of the bodies is 10×10 [mm], the top face of the lower body Ω^1 and the bottom face of the upper body Ω_2 are described by the functions ϕ_1 and ϕ_2, respectively, where

$$\phi_1(x, y) = 5 \sin\left(20 - \sqrt{400 - x^2 - y^2}\right), \qquad \phi_2(x, y) = \sin\left(\sqrt{20.2^2 - x^2 - y^2} - 20.2\right).$$

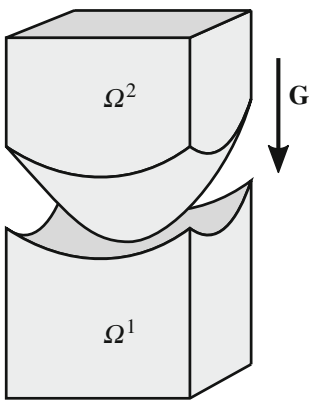

Fig. 13.2 Geometry with traces of domain decomposition

Material constants are defined by the Young modulus $E^1 = E^2 = 2.1 \cdot 10^5$ [MPa], the Poisson ratio $\nu^1 = \nu^2 = 0.3$, and the density $\rho^1 = \rho^2 = 7.85 \cdot 10^{-3}$ [g/mm^3]. The initial gap between the curved boxes is set to $g_0 = 0.001$ [mm]. The initial

velocity of the upper body Ω^2 in the vertical direction is $v^2 = -1$ [m/s]. The upper body is floating and the lower body is fixed along the bottom side. The linearized non-interpenetration condition was imposed on the contact interface.

The space discretization was carried out with the brick elements using varying discretization and decomposition parameters h and H, respectively. We kept $H/h = 16$. The number of subdomains ranged from 16 to 250, the number of dual variables ranged from 21,706 to 443,930, and the number of primal variables ranged from 196,608 to 3,072,000. For the time discretization, we used the contact-stabilized Newmark algorithm with the constant time step $\Delta = 3h10^{-3}$ and solved the impact of bodies in the time $t = [0, 45\Delta]$.

The energy development is in Fig. 13.3 shows that the algorithm preserves the energy as predicted by the theory. The development of contact pressure in the middle of Γ_C^2 is in Fig. 13.4.

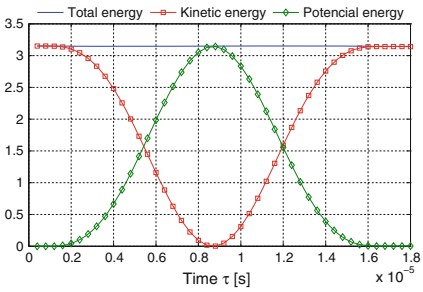

Fig. 13.3 Energy conservation [ton · mm^2 · s^{-2}]

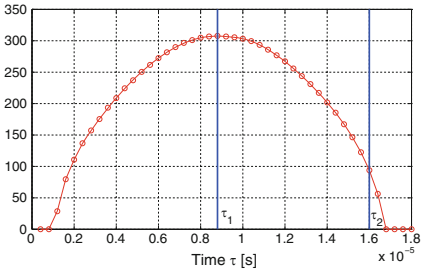

Fig. 13.4 Contact pressure in the middle of Γ_C^2 [MPa]

The performance of the algorithm is in Fig. 13.5.

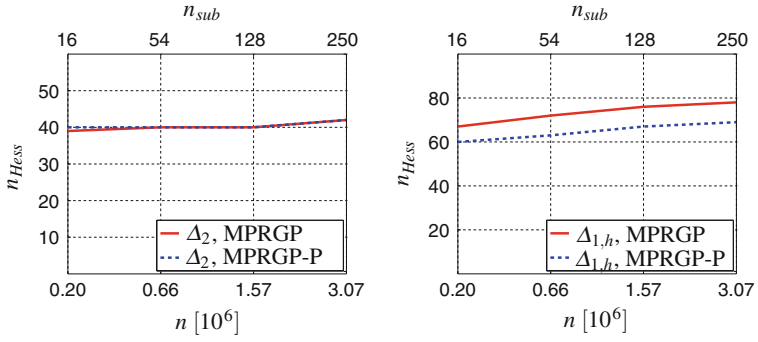

Fig. 13.5 Hessian multiplications for MPRGP and MPRGP-P, Δ_2 (*left*) and $\Delta_{1,h}$ (*right*)

13.8.2 Impact of Three Bodies

We considered the transient analysis of three elastic bodies in mutual contact. The bottle is a sample model of Autodesk Inventor ("bottle.ipt"). We prescribed the initial velocity $v^3 = -5$ [m/s] on the sphere in the horizontal direction. The L-shape body was fixed along the bottom side. For the time discretization, we used the constant time step $\Delta = 1 \cdot 10^{-3}$ [s] and solved the impact of bodies in the time interval $t = [0, 150\Delta]$ [s].

The solution (total displacement of the problem discretized by $1.2 \cdot 10^5$ primal and $8.5 \cdot 10^3$ dual variables and decomposed into 32 subdomains using METIS in time $t_1 = 20\Delta$, $t_2 = 40\Delta$, $t_3 = 60\Delta$, $t_4 = 80\Delta$, $t_5 = 100\Delta$, $t_6 = 120\Delta$ [s]) is depicted in Fig. 13.6. The whole computational process required approximately 300 and 30 matrix–vector multiplications per time step during the impact and the separation, respectively.

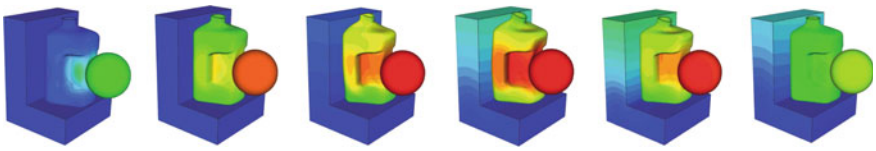

Fig. 13.6 Impact of bodies in time

13.9 Comments

Convenient presentation of the variational formulation, the finite element approximation of the solution, and standard iterative methods for the solution can be found in the book by Kikuchi and Oden [16]. More on the formulation of continuous problem and the existence of solutions can be found in Eck, Jarušek, and Krbec [17]. An up-to-date engineering approach to the solution of transient contact problems can be found in Laursen [18] or Wriggers [19]. See Chap. 15 for the discussion of the combination of TFETI and the mortar approximation of contact conditions.

A conveniently applicable stable energy conserving method, which is based on quadrature formulas and applicable to frictional contact problems, was proposed by Hager, Hüber, and Wohlmuth [20]. The method is based on quadrature formulas. For its analysis, see Hager and Wohlmuth [21]. Important issues related to the time discretization and avoiding the nonphysical oscillations that result from the application of standard time discretization methods for unconstrained problems were proposed, e.g., by Chawla and Laursen [22], Wohlmuth [2], Bajer and Demkowicz [23, 24], Khenous, Laborde, and Renard [25, 26], Deuflhard, Krause, and Ertel [27], and Kornhuber et al. [28]. A discussion of the advantages and disadvantages of the respective approaches can be found in [29].

The experimental evidence of optimal complexity of the time step for some academic problems has been reported by the methods discussed in the previous chapters, in particular in Chap. 11. Thus Krause and Walloth [4] documented that even some transient problems with friction can be solved very efficiently by the monotone multigrid method. See also Wohlmuth and Krause [30] and Kornhuber et al. [28]. As mentioned above, the multigrid methods typically reduce the solution of auxiliary problems to a sequence of linear problems that are solved with optimal complexity, typically leaving the nonlinear steps without theoretically supported preconditioning, so that the theory does not guarantee the optimal performance of the overall procedure.

The combination of the standard finite element space discretization with the contact-stabilized Newmark scheme that we use here was introduced by Krause and Walloth [4] for the transient problems with friction. The results presented in this chapter appeared in Dostál et al. [15]. The preconditioning by conjugate projector (deflation) for linear problems was introduced independently by Marchuk and Kuznetsov [31], Nicolaides [32], and Dostál [12]). For recent references see Gutnecht [33]. The procedure was adapted to the solution of inequality constrained problems by Domorádová–Jarošová and Dostál [9]. The lack of sufficiently small subspace with the solution does not allow stronger theoretical results. Farhat, Chen, and Mandel [8] used the preconditioning by the conjugate projector onto the subspace defined by the natural coarse grid to achieve optimality of FETI for unconstrained transient problems.

References

1. Zhao, X., Li, Z.: The solution of frictional wheel-rail rolling contact with a 3D transient finite element model: validation and error analysis. Wear **271**(1–2), 444–452 (2011)
2. Wohlmuth, B.I.: Variationally consistent discretization scheme and numerical algorithms for contact problems. Acta Numerica **20**, 569–734 (2011)
3. Jarošová, M., Klawonn, A., Rheinbach, O.: Projector preconditioning and transformation of basis in FETI-DP algorithms for contact problems. Math. Comput. Simul. **82**(10), 1894–1907 (2012)
4. Krause, R., Walloth, M.: A time discretization scheme based on Rothe's method for dynamical contact problems with friction. Comput. Methods Appl. Mech. Eng. **199**, 1–19 (2009)
5. Brenner, S.C.: The condition number of the Schur complement. Numerische Mathematik **83**, 187–203 (1999)
6. Wathen, A.J.: Realistic eigenvalue bounds for the Galerkin Mass Matrix. IMA J. Numer. Anal. **7**, 449–457 (1987)
7. Toselli, A., Widlund, O.B.: Domain Decomposition Methods – Algorithms and Theory. Springer Series on Computational Mathematics, vol. 34. Springer, Berlin (2005)
8. Farhat, C., Chen, J., Mandel, J.: A scalable Lagrange multiplier based domain decomposition method for time-dependent problems. Int. J. Numer. Methods Eng. **38**, 3831–3853 (1995)
9. Domorádová, M., Dostál, Z.: Projector preconditioning for partially bound constrained quadratic optimization. Numer. Linear Algebra Appl. **14**(10), 791–806 (2007)
10. Farhat, C., Mandel, J., Roux, F.-X.: Optimal convergence properties of the FETI domain decomposition method. Comput. Methods Appl. Mech. Eng. **115**, 365–385 (1994)

11. Iontcheva, A.H., Vassilevski, P.S.: Monotone multigrid methods based on element agglomeration coarsening away from the contact boundary for the Signorini's problem. Numer. Linear Algebra Appl. **11**(2–3), 189–204 (2004)
12. Dostál, Z.: Conjugate gradient method with preconditioning by projector. Int. J. Comput. Math. **23**, 315–323 (1988)
13. Dostál, Z.: Optimal Quadratic Programming Algorithms, with Applications to Variational Inequalities, 1st edn. Springer, New York (2009)
14. Kozubek, T., Markopoulos, A., Brzobohatý, T., Kučera, R., Vondrák, V., Dostál, Z.: MatSol–MATLAB efficient solvers for problems in engineering. http://industry.it4i.cz/en/products/matsol/ (2015)
15. Dostál, Z., Kozubek, T., Brzobohatý, T., Markopoulos, A., Vlach, O.: Scalable TFETI with optional preconditioning by conjugate projector for transient contact problems of elasticity. Comput. Methods Appl. Mech. Eng. **247–248**, 37–50 (2012)
16. Kikuchi, N., Oden, J.T.: Contact Problems in Elasticity. SIAM, Philadelphia (1988)
17. Eck, C., Jarůšek, J., Krbec, M.: Unilateral Contact Problems. Chapman & Hall/CRC, London (2005)
18. Laursen, T.: Computational Contact and Impact Mechanics. Springer, Berlin (2002)
19. Wriggers, P.: Contact Mechanics. Springer, Berlin (2005)
20. Hager, C., Hüber, S., Wohlmuth, B.I.: A stable energy conserving approach for frictional contact problems based on quadrature formulas. Int. J. Numer. Methods Eng. **73**, 205–225 (2008)
21. Hager, C., Wohlmuth, B.I.: Analysis of a space-time discretization for dynamic elasticity problems based on mass-free surface elements. SIAM J. Numer. Anal. **47**(3), 1863–1885 (2009)
22. Chawla, V., Laursen, T.: Energy consistent algorithms for frictional contact problems. Int. J. Numer. Methods Eng. **42**, 799–827 (1998)
23. Bajer, A., Demkowicz, L.: Conservative discretization of contact/impact problems for nearly rigid bodies. Comput. Methods Appl. Mech. Eng. **190**, 1903–1924 (2001)
24. Bajer, A., Demkowicz, L.: Dynamic contact/impact problems, energy conservation, and planetary gear trains. Comput. Methods Appl. Mech. Eng. **191**, 4159–4191 (2002)
25. Khenous, H., Laborde, P., Renard, Y.: On the discretization of contact problems in elastodynamics. Analysis and Simulation of Contact Problems. Lecture Notes in Applied and Computational Mechanics, vol. 27, pp. 31–38. Springer, Berlin (2006)
26. Khenous, H., Laborde, P., Renard, Y.: Mass redistribution method for finite element contact problems in elastodynamics. Eur. J. Mech. A/Solids **27**, 918–932 (2008)
27. Deuflhard, P., Krause, R., Ertel, S.: A contact-stabilized newmark method for dynamical contact problems. Int. J. Numer. Methods Eng. **73**(9), 1274–1290 (2008)
28. Kornhuber, R., Krause, R., Sander, O., Deuflhard, P., Ertel, S.: A monotone multigrid solver for two body contact problems in biomechanics. Comput. Vis. Sci. **11**, 3–15 (2008)
29. Krause, R., Walloth, M.: Presentation and comparison of selected algorithms for dynamic contact based on the Newmark scheme. Appl. Numer. Math. **62**(10), 1393–1410 (2012)
30. Wohlmuth, B.I., Krause, R.: Monotone methods on nonmatching grids for nonlinear contact problems. SIAM J. Sci. Comput. **25**, 324–347 (2003)
31. Marchuk, G.I.: Kuznetsov, YuA: theory and applications of the generalized conjugate gradient method. Adv. Math. Suppl. Stud. **10**, 153–167 (1986)
32. Nicolaides, R.A.: Deflation of conjugate gradients with applications to boundary value problems. SIAM J. Numer. Anal. **24**, 355–365 (1987)
33. Gutnecht, M.H.: Spectral deflation in Krylov solvers: a theory of coordinate based methods. Electron. Trans. Numer. Anal. **39**, 156–185 (2012)

Chapter 14
TBETI

The optimality results presented in the previous four chapters used the formulation of the conditions of equilibrium in the domains which were decomposed into nonoverlapping subdomains and "glued" by the equality constraints. The subdomains were discretized by means of the finite element method. Using the duality, the original problem was reduced to the minimization of a quadratic function in Lagrange multipliers with a well-conditioned Hessian subject to some separable inequality constraints and linear equality constraints. Since the multipliers were associated with the nodes on the boundaries of the subdomains, the procedure can be viewed also as the reduction to the *skeleton of the decomposition*.

There is a natural question whether we can reduce the formulation of the conditions of equilibrium to the boundary on continuous level. Such reduction promises a lower dimension of the primal discretized problem, simplified discretization, higher precision of numerical solution, improved flexibility in solving the problems that include a free boundary as in the contact shape optimization, and an effective treatment of the boundary value problems defined on unbounded domains.

In this chapter, we shall indicate how to reduce frictionless contact problems to the skeleton of the decomposition. The basic observation is that it is possible to find the formulae for sufficiently rich sets of so-called fundamental solutions that can be combined so that they satisfy both the boundary conditions of the original problem and the "gluing" conditions on the artificial interface.

Our development of scalable algorithms for contact problems is based on the BETI (Boundary Element Tearing and Interconnecting) method, which was introduced by Langer and Steinbach [1]. The main new feature of BETI is that it generates the counterparts of the Schur complements directly by the discretization of the conditions of equilibrium reduced to the boundary. We use a variant of BETI called TBETI, which is the boundary element counterpart of TFETI. Here we describe the main steps of reducing the contact problems to the boundary—first for scalar problems, then for elasticity.

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3_14

14.1 Green's Representation Formula for 2D Laplace Operator

To explain the fundamental ideas of boundary element methods, let us first consider the scalar problems governed by 2D Laplace's operator introduced in Chap. 10. Let $\Omega \subset \mathbb{R}^2$ be a bounded Lipschitz domain with the boundary Γ, such as Ω^{ij} in the benchmarks of Chap. 10, but we shall assume that the diameter of $\overline{\Omega}$ satisfies

$$\operatorname{diam} \overline{\Omega} = \max_{\mathbf{x},\mathbf{y}\in\overline{\Omega}} \|\mathbf{x}-\mathbf{y}\| < 1.$$

Though our purpose here is only to present the basic ideas in the extent that is necessary for understating the algorithms, we shall use some abstract concepts to simplify the formulation of true statements and identify the steps that are relevant for implementation. In particular, we shall use the standard interior trace operator γ_0 (see (4.3)) and the associated interior conormal derivative operator γ_1 (see (4.6)).

If $u \in C^\infty(\overline{\Omega})$, then

$$\gamma_0 u = u|\Gamma, \qquad \langle \gamma_1 u, v\rangle = \int_\Gamma \frac{\partial u}{\partial \mathbf{n}} v \,\mathrm{d}\Gamma, \quad v \in C^\infty\left(\mathbb{R}^2\right),$$

where $\mathbf{n} = \mathbf{n}(\mathbf{x})$ is the outer unit normal vector to Γ at $\mathbf{x}$. In this case, we can identify

$$\gamma_1 u = \frac{\partial u}{\partial \mathbf{n}} = \nabla u \cdot \mathbf{n}, \quad \mathbf{x} \in \Gamma.$$

Let

$$U(\mathbf{x},\mathbf{y}) = -\frac{1}{2\pi} \log \|\mathbf{x}-\mathbf{y}\|$$

denote the *fundamental solution* of Δ in $\mathbb{R}^2$. If $\delta_\mathbf{y}$ denotes Dirac's δ-distribution at $\mathbf{y}$, then

$$-\Delta_\mathbf{x} U(\mathbf{x},\mathbf{y}) = \delta_\mathbf{y} \quad \text{for } \mathbf{y} \in \mathbb{R}^2$$

in the sense of distributions in $\mathbb{R}^2$, i.e.,

$$-\int_{\mathbb{R}^2} \Delta U(\mathbf{x},\mathbf{y}) v(\mathbf{x}) \,\mathrm{d}\Omega_\mathbf{x} = v(\mathbf{y})$$

for any $\mathbf{y} \in \mathbb{R}^2$ and $v \in C^\infty(\mathbb{R}^2)$ with a compact support. A key tool on the way to a boundary formulation is the following theorem.

Theorem 14.1 (Green's Representation Formula) *Let $u \in H^1_\Delta(\Omega)$ be such that $-\Delta u = f$ on Ω. Then for any $\mathbf{x} \in \Omega$*

$$u(\mathbf{x}) = \int_{\Omega} f(\mathbf{y})\, U(\mathbf{x},\mathbf{y})\, \mathrm{d}\Omega_{\mathbf{y}} + \int_{\Gamma} \gamma_1 u(\mathbf{y}) U(\mathbf{x},\mathbf{y})\, \mathrm{d}\Gamma_{\mathbf{y}} - \int_{\Gamma} \gamma_0 u(\mathbf{y})\, \frac{\partial}{\partial \mathbf{n}_{\mathbf{y}}} U(\mathbf{x},\mathbf{y})\, \mathrm{d}\Gamma_{\mathbf{y}}. \tag{14.1}$$

Proof See, e.g., [2]. □

Since all three summands of (14.1) belong to H^1_{Δ}, we can apply the operators γ_0, γ_1 to the representation formula (14.1) to get

$$\begin{aligned}
\frac{1}{2}\gamma_0 u(\mathbf{x}) &= \int_{\Omega} f(\mathbf{y})\, U(\mathbf{x},\mathbf{y})\, \mathrm{d}\Omega_{\mathbf{y}} + \int_{\Gamma} \gamma_1 u(\mathbf{y})\, U(\mathbf{x},\mathbf{y})\, \mathrm{d}\Gamma_{\mathbf{y}} \\
&\qquad - \int_{\Gamma} \gamma_0 u(\mathbf{y})\, \frac{\partial}{\partial \mathbf{n}_{\mathbf{y}}} U(\mathbf{x},\mathbf{y})\, \mathrm{d}\Gamma_{\mathbf{y}}, \\
\frac{1}{2}\gamma_1 u(\mathbf{x}) &= \gamma_{1,\mathbf{x}} \int_{\Omega} f(\mathbf{y}) U(\mathbf{x},\mathbf{y})\, \mathrm{d}\Omega_{\mathbf{y}} + \int_{\Gamma} \gamma_1 u(\mathbf{y})\, \frac{\partial}{\partial \mathbf{n}_{\mathbf{x}}} U(\mathbf{x},\mathbf{y})\, \mathrm{d}\Gamma_{\mathbf{y}} \\
&\qquad - \gamma_{1,\mathbf{x}} \int_{\Gamma} \gamma_0 u(\mathbf{y})\, \frac{\partial}{\partial \mathbf{n}_{\mathbf{y}}} U(\mathbf{x},\mathbf{y})\, \mathrm{d}\Gamma_{\mathbf{y}}.
\end{aligned}$$

After introducing the standard operators (see [3–5]), we get Calderon's system of integral equations valid on Γ

$$\begin{bmatrix} \gamma_0 u \\ \gamma_1 u \end{bmatrix} = \begin{bmatrix} \frac{1}{2} I - K & V \\ D & \frac{1}{2} I + K' \end{bmatrix} \begin{bmatrix} \gamma_0 u \\ \gamma_1 u \end{bmatrix} + \begin{bmatrix} N_0 f \\ N_1 f \end{bmatrix} \tag{14.2}$$

with the following integral operators defined for $\mathbf{x} \in \Gamma$:

single-layer potential operator

$$(Vt)(\mathbf{x}) = \int_{\Gamma} t(\mathbf{y})\, U(\mathbf{x},\mathbf{y})\, \mathrm{d}\Gamma_{\mathbf{y}}, \quad V : H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma),$$

double-layer potential operator

$$(Kh)(\mathbf{x}) = \int_{\Gamma} h(\mathbf{y})\, \frac{\partial}{\partial \mathbf{n}_{\mathbf{y}}} U(\mathbf{x},\mathbf{y})\, \mathrm{d}\Gamma_{\mathbf{y}}, \quad K : H^{1/2}(\Gamma) \to H^{1/2}(\Gamma),$$

adjoint double-layer potential operator

$$(K't)(\mathbf{x}) = \int_{\Gamma} t(\mathbf{y})\, \frac{\partial}{\partial \mathbf{n}_{\mathbf{x}}} U(\mathbf{x},\mathbf{y})\, \mathrm{d}\Gamma_{\mathbf{y}}, \quad K' : H^{-1/2}(\Gamma) \to H^{-1/2}(\Gamma),$$

hypersingular integral operator

$$(Dh)(\mathbf{x}) = -\gamma_{1,\mathbf{x}} \int_{\Gamma} h(\mathbf{y})\, \frac{\partial}{\partial \mathbf{n}_{\mathbf{y}}} U(\mathbf{x},\mathbf{y})\, \mathrm{d}\Gamma_{\mathbf{y}}, \quad D : H^{1/2}(\Gamma) \to H^{-1/2}(\Gamma),$$

and *Newton's potential operators*

$$(N_0 f)(\mathbf{x}) = \int_\Omega f(\mathbf{y})\, U(\mathbf{x},\mathbf{y})\, \mathrm{d}\Omega_\mathbf{y}, \quad N_0 : L^2(\Omega) \to H^{1/2}(\Gamma),$$
$$(N_1 f)(\mathbf{x}) = \gamma_{1,\mathbf{x}} \int_\Omega f(\mathbf{y})\, U(\mathbf{x},\mathbf{y})\, \mathrm{d}\Omega_\mathbf{y}, \quad N_1 : L^2(\Omega) \to H^{-1/2}(\Gamma).$$

The mapping properties of the above integral operators are well known, in particular all are bounded, V is symmetric and elliptic (see, e.g., Steinbach [4], Theorem 6.23), and D is symmetric and semi-elliptic with the kernel R of the solutions of the Neumann homogeneous boundary value problem

$$\Delta u = 0 \quad \text{in } \Omega,$$
$$\gamma_1 u = 0 \quad \text{on } \Gamma.$$

14.2 Steklov–Poincaré Operator

For a given $f \in L^2(\Omega)$, observing that the assumption on the diameter of Ω implies that the operator V is elliptic, we get from the first equation of (14.2) the Dirichlet–Neumann map

$$\gamma_1 u = V^{-1}\left(\frac{1}{2}I + K\right)\gamma_0 u - V^{-1} N_0 f \quad \text{on } \Gamma. \tag{14.3}$$

We can also substitute the latter formula into the second equation of (14.2) to get another representation of the Dirichlet–Neumann map

$$\gamma_1 u = \left(\left(\frac{1}{2}I + K'\right)V^{-1}\left(\frac{1}{2}I + K\right) + D\right)\gamma_0 u + \left(N_1 - \left(\frac{1}{2}I + K'\right)V^{-1}N_0\right) f \tag{14.4}$$

on Γ. The last equation can be expressed by means of two new operators, the *Steklov–Poincaré operator* defined by the equivalent representations

$$S = V^{-1}\left(\frac{1}{2}I + K\right) \tag{14.5}$$
$$= \left(\frac{1}{2}I + K'\right)V^{-1}\left(\frac{1}{2}I + K\right) + D \tag{14.6}$$

and the *Newton operator* defined by

$$N = V^{-1}N_0 = \left(N_1 - \left(\frac{1}{2}I + K'\right)V^{-1}N_0\right).$$

The Eqs. (14.3) and (14.4) now read

$$\gamma_1 u = S\gamma_0 u - Nf \quad \text{on} \quad \Gamma. \tag{14.7}$$

If $g : \Gamma \to \mathbb{R}$ is sufficiently smooth, then the solution of

$$-\Delta u = f, \quad \gamma_0 u = g \tag{14.8}$$

can be obtained in the form

$$u = u_g + u_f,$$

where u_g, u_f solve

$$\begin{aligned} -\Delta u_g = 0, \; -\Delta u_f = f \quad &\text{in} \quad \Omega, \\ \gamma_0 u_g = g, \quad \gamma_0 u_f = 0 \quad &\text{on} \quad \Gamma. \end{aligned} \tag{14.9}$$

Moreover, (14.7) implies that for any $v \in H^{1/2}(\Gamma)$

$$\langle Sg, v\rangle = \langle \gamma_1 u_g, v\rangle = \int_\Gamma \frac{\partial}{\partial \mathbf{n}} u_g(\mathbf{x}) v(\mathbf{x}) \, \mathrm{d}\Gamma \tag{14.10}$$

and

$$\langle Nf, v\rangle = -\langle \gamma_1 u_f, v\rangle = -\int_\Gamma \frac{\partial}{\partial \mathbf{n}} u_f(\mathbf{x}) v(\mathbf{x}) \, \mathrm{d}\Gamma. \tag{14.11}$$

The properties of the Steklov–Poincaré operator are summarized in the following theorem.

Theorem 14.2 *The Steklov–Poincaré operator*

$$S : \; H^{1/2}(\Gamma) \to H^{-1/2}(\Gamma)$$

is symmetric and there are $\alpha^D > 0$ *and* $\alpha > 0$ *such that*

$$\langle Sv, v\rangle \geq \alpha^D \, \|v\|^2_{H^{1/2}(\Gamma)} \quad \text{for all } v \in H^{1/2}(\Gamma)/R$$

and

$$\langle Sv, v\rangle \geq \alpha \, \|v\|^2_{H^{1/2}(\Gamma)} \quad \text{for all } v \in H_0^{1/2}(\Gamma, \Gamma_U),$$

where

$$H_0^{1/2}(\Gamma, \Gamma_U) = \{v \in H^{1/2}(\Gamma) : \; v = 0 \;\text{ on } \Gamma_U\}$$

and Γ_U *is a part of the boundary* Γ *with a positive measure.*

Proof See the books by McLean [2] od Steinbach [4, 6]. □

Note that the representation (14.5) together with a Galerkin discretization typically results in nonsymmetric stiffness matrices. The symmetry of the stiffness matrix is essential in our further analysis, so we consider only the symmetric representation (14.6).

14.3 Decomposed Boundary Variational Inequality

Now we are ready to reduce the scalar variational inequality that was introduced in Sect. 10.3 to the boundaries of the subdomains. Let us recall that a solution of the variational inequality (10.10) defined on the union of the subdomains Ω^{ij} requires to find $u \in \mathscr{K}_{DD}$ such that

$$a(u, v-u) \geq \ell(v-u), \quad v \in \mathscr{K}_{DD}, \tag{14.12}$$

where

$$\begin{aligned}
V^{ij} &= \left\{ v^{ij} \in H^1(\Omega^{ij}) : \gamma_0 v^{ij} = 0 \ \text{ on } \ \Gamma_U^i \cap \overline{\Omega}^{ij} \right\}, \quad i = 1, 2, \quad j = 1, \ldots, p, \\
V_{DD} &= (V^{11} \times \cdots \times V^{1p}) \times (V^{21} \times \cdots \times V^{2p}), \\
\mathscr{K}_{DD} &= \left\{ v \in V_{DD} : \gamma_0 v^{2i} \geq \gamma_0 v^{1j} \text{ on } \Gamma_C^{2i} \cap \Gamma_C^{1j} \text{ and } \gamma_0 v^{ij} = \gamma_0 v^{ik} \ \text{ on } \ \Gamma^{ij} \cap \Gamma^{ik} \right\},
\end{aligned}$$

and

$$\begin{aligned}
(u, v) &= \sum_{i=1}^{2} \sum_{j=1}^{p} \int_{\Omega^{ij}} u^{ij} v^{ij} \, \mathrm{d}\Omega, \\
a(u, v) &= \sum_{i=1}^{2} \sum_{j=1}^{p} \int_{\Omega^{ij}} \left(\frac{\partial u^{ij}}{\partial x_1} \frac{\partial v^{ij}}{\partial x_1} + \frac{\partial u^{ij}}{\partial x_2} \frac{\partial v^{ij}}{\partial x_2} \right) \mathrm{d}\Omega, \\
\ell(v) &= (f, v).
\end{aligned} \tag{14.13}$$

If u is a sufficiently smooth solution of (14.12), then u^i defined by

$$u^i = u^{ij} \quad \text{for} \quad \mathbf{x} \in \overline{\Omega}^{ij}, \quad i = 1, 2, \ j = 1, \ldots, p,$$

is a classical solution of (10.1)–(10.2).

To reduce the variational formulation of the decomposed problem to the skeleton

$$\Sigma = \Gamma^{11} \times \cdots \times \Gamma^{1p} \times \Gamma^{21} \times \cdots \times \Gamma^{2p},$$

let

$$V_b^{ij} = \left\{ v^{ij} \in H^{1/2}(\Gamma^{ij}) : \ v^{ij} = 0 \ \text{ on } \ \Gamma_U^i \cap \Gamma^{ij} \right\}, \quad i = 1, 2, \quad j = 1, \ldots p,$$

denote the closed subspaces of $H^{1/2}(\Gamma^{ij})$, and let

$$V_{DD}^b = (V_b^{11} \times \cdots \times V_b^{1p}) \times (V_b^{21} \times \cdots \times V_b^{2p}),$$
$$\mathscr{K}_{DD}^b = \left\{ v \in V_{DD}^b : v^{2i} - v^{1j} \geq 0 \text{ on } \Gamma_C^{2i} \cap \Gamma_C^{1j} \text{ and } v^{ij} = v^{ik} \text{ on } \Gamma^{ij} \cap \Gamma^{ik} \right\}.$$

On V_{DD}^b we shall define a scalar product

$$(u, v) = \sum_{i=1}^{2} \sum_{j=1}^{p} \int_{\Gamma^{ij}} u^{ij} v^{ij} \, \mathrm{d}\Gamma,$$

a symmetric bilinear form

$$a^b(u, v) = \sum_{i=1}^{2} \sum_{j=1}^{p} \langle S^{ij} u^{ij}, v^{ij} \rangle,$$

and a linear form

$$\ell^b(v) = \sum_{i=1}^{2} \sum_{j=1}^{p} \langle N^{ij} f^{ij}, v^{ij} \rangle,$$

where the ordered couples of upper indices ij associate the objects with the subdomains Ω^{ij}.

Let us now assume that $u \in \mathscr{K}_{DD}$ is a solution of (14.12) and $v \in \mathscr{K}_{DD}$. After substituting $\Delta u = -f$ into Green's identity (see Theorem 4.2)

$$\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}\Omega + \int_\Omega \Delta u \, v \, \mathrm{d}\Omega = \langle \gamma_1 u, \gamma_0 v \rangle$$

and using (14.7), (14.10), and (14.11) to its right-hand side, we get

$$a(u, v - u) - \ell(v - u) = a^b \left(\gamma_0 u, \gamma_0 (v - u) \right) - \ell^b \left(\gamma_0 (v - u) \right). \tag{14.14}$$

Thus if $u \in \mathscr{K}_{DD}$ is a solution of (14.12), then their traces on Γ satisfy

$$a^b \left(\gamma_0 u, \gamma_0 (v - u) \right) \geq \ell^b \left(\gamma_0 (v - u) \right), \quad v \in \mathscr{K}_{DD}.$$

Moreover, if we define

$$q^b : V_{DD}^b \to \mathbb{R}, \quad q^b(v) = \frac{1}{2} a^b(v, v) - \ell^b(v),$$

we can use the same arguments as in Sect. 10.3 to get that any solution $u \in \mathcal{K}_{DD}$ of (14.12) satisfies

$$q^b(\gamma_0 u) \le q^b(v), \quad v \in \mathcal{K}^b_{DD}. \tag{14.15}$$

Now we are ready to formulate the main result of this section.

Theorem 14.3 *Let $g \in \mathcal{K}^b_{DD}$ satisfy*

$$q^b(g) \le q^b(v), \quad v \in \mathcal{K}^b_{DD}, \tag{14.16}$$

and let $u = u_g + u_f$ denote a solution of (14.9) with u_f, u_g defined by (14.9). Then u solves (14.12).

Proof The proofs follow the definitions and (14.14). □

14.4 Boundary Discretization and TBETI

The boundary element discretization of the variational inequality defined on the skeleton Σ of the decomposition (10.11) is very similar to the finite element discretization defined on Ω. The only difference is that the basis functions ϕ^ℓ are defined on the boundaries of the subdomains Ω^{ij}.

Let us assume that each domain Ω^i is decomposed into $p = 1/H^2$ square subdomains Ω^{ij}, $i = 1, 2$, $j = 1, \dots, p$, as in Sect. 10.4, and decompose the boundary Γ^{ij} of Ω^{ij} into n^τ_{ij} line segments τ^{ij}_ℓ of the length h. On each τ^{ij}_ℓ, we can choose one or more nodes and use them to define the shape functions (see, e.g., Gaul et al. [7], Sect. 4.2.5) or the boundary element polynomial basis functions. For example, we can use the piecewise constant functions Ψ^{ij}_ℓ associated with the elements τ^{ij}_ℓ or the linear basis functions ϕ^{ij}_ℓ associated with the vertices of the elements which are depicted in Fig. 14.1. We denote the number of the basis functions by n_{ij}.

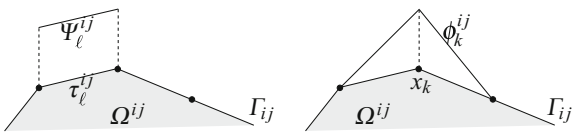

Fig. 14.1 Piecewise constant (*left*) and piecewise linear continuous (*right*) basis function

We shall look for an approximate solution u_h in the trial space V_h which is spanned by the basis functions,

$$V_h = V_h^{11} \times \cdots \times V_h^{1p} \times V_h^{21} \times \cdots \times V_h^{2p},$$
$$V_h^{ij} = \mathrm{Span}\{\phi^{ij}_\ell, \ell = 1, \dots, n_{ij}\}.$$

Decomposing u_h into components, i.e.,

$$u_h = (u_h^{11}, \ldots, u_h^{1p}, u_h^{21}, \ldots, u_h^{2p}),$$
$$u_h^{ij}(\mathbf{x}) = \sum_{\ell=1,\ldots,n_{ij}} u_\ell^{ij} \phi_\ell^{ij}(\mathbf{x}),$$

we get

$$\langle S u_h, u_h \rangle = \sum_{i=1,2;\ j=1,\ldots,p} \langle S^{ij} u_h^{ij}, u_h^{ij} \rangle,$$
$$\langle S^{ij} u_h^{ij}, u_h^{ij} \rangle = \sum_{\ell,m=1,\ldots,n_{ij}} u_\ell^{ij} \langle S^{ij} \phi_\ell^{ij}, \phi_m^{ij} \rangle u_m^{ij} = (\mathbf{u}^{ij})^T \mathsf{S}^{ij} \mathbf{u}^{ij},$$
$$\left[\mathsf{S}^{ij}\right]_{\ell m} = \langle S^{ij} \phi_\ell^{ij}, \phi_m^{ij} \rangle, \quad [\mathbf{u}^{ij}]_\ell = u_\ell^{ij}.$$

Similarly

$$\ell^b(u_h) = \sum_{i=1}^{2} \sum_{j=1}^{p} \langle N^{ij} f^{ij}, u_h^{ij} \rangle,$$
$$\langle N^{ij} f^{ij}, u_h^{ij} \rangle = \sum_{\ell=1,\ldots,n_{ij}} \langle N^{ij} f^{ij}, u_\ell^{ij} \phi_\ell^{ij} \rangle = (\mathbf{f}_{ij}^{\,b})^T \mathbf{u}^{ij},$$
$$\left[\mathbf{f}_{ij}^{\,b}\right]_\ell = \langle N^{ij} f^{ij}, \phi_\ell^{ij} \rangle. \tag{14.17}$$

The effective evaluation of (14.17) can be found, e.g., in Steinbach [4], Chap. 10.

Let us now assign each subdomain a unique index and let us index contiguously the nodes and the entries of corresponding vectors in subdomains, so that we can write

$$\mathsf{S}^b = \begin{bmatrix} \mathsf{S}_1^b & \mathsf{O} & \cdots & \mathsf{O} \\ \mathsf{O} & \mathsf{S}_2^b & \cdots & \mathsf{O} \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{O} & \mathsf{O} & \cdots & \mathsf{S}_s^b \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_s \end{bmatrix}, \quad \mathbf{f}^b = \begin{bmatrix} \mathbf{f}_1^{\,b} \\ \vdots \\ \mathbf{f}_s^{\,b} \end{bmatrix}.$$

The SPS matrix $\mathsf{S}^b \in \mathbb{R}^{n \times n}$ denotes a discrete analog of the Steklov–Poincaré operator that assigns each vector $\mathbf{u}$ of nodal boundary displacements the vector of corresponding nodal forces. Our notation indicates that S^b is closely related to the Schur complement S introduced in Chap. 10. If we denote by B_I^b and B_E^b the full rank matrices which describe the discretized non-penetration and gluing conditions,

respectively, we get the discretized version of problem (10.11) with auxiliary domain decomposition that reads

$$\min \frac{1}{2}\mathbf{u}^T \mathsf{S}^b \mathbf{u} - (\mathbf{f}^b)^T \mathbf{u} \quad \text{s.t.} \quad \mathsf{B}_I^b \mathbf{u} \le \mathbf{o} \quad \text{and} \quad \mathsf{B}_E^b \mathbf{u} = \mathbf{o}. \tag{14.18}$$

In (14.18), the matrices B_I^b and B_E^b can be obtained from the matrices B_E and B_I used in (10.12) by deleting the columns which correspond to the inner nodes of the subdomains. In particular, we can easily achieve

$$\mathsf{B}^b \left(\mathsf{B}^b\right)^T = \mathsf{I}, \quad \mathsf{B}^b = \begin{bmatrix} \mathsf{B}_I^b \\ \mathsf{B}_E^b \end{bmatrix}.$$

The next steps are very similar to those described in Chap. 10 (details can be found also in [8]), so we shall switch to the main goal of this chapter, the development of a scalable BETI-based algorithm for solving the frictionless contact problems of elasticity.

14.5 Operators of Elasticity

To extend our exposition to frictionless contact problems in 3D, let us briefly recall some fundamental results concerning the boundary integral operators that can be used to reduce the conditions of equilibrium of a homogeneous elastic body to its boundary. Though most of these results are rather complicated, they can be obtained by the procedures indicated in Sect. 14.1 with Green's formulae replaced by Betti's formulae. The most important point is that the energy associated with a sufficiently smooth displacement $\mathbf{v}$ which satisfies the conditions of equilibrium in the domain can be evaluated by means of the values of $\mathbf{v}$ and its derivatives on the boundary as in our scalar benchmark (14.14). More details can be found, e.g., in the paper by Costabel [3] or in the books by Steinbach [4], Rjasanow and Steinbach [5], or McLean [2]. See also Dostál et al. [9].

Let $\Omega \subset \mathbb{R}^3$ be a bounded Lipschitz domain with the boundary Γ which is filled with a homogeneous isotropic material and consider the elliptic operator $\mathscr{L}$ which assigns each sufficiently smooth $\mathbf{u} : \Omega \to \mathbb{R}^3$ a mapping $\mathscr{L}\mathbf{u} : \Omega \to \mathbb{R}^3$ by

$$\mathscr{L}\mathbf{u} = -\operatorname{div} \sigma(\mathbf{u}) \quad \text{in} \quad \Omega,$$

where the stress tensor σ satisfies, as in Sect. 11.2, Hooke's law

$$\sigma(\mathbf{v}) = \mathsf{C}\varepsilon(\mathbf{v}).$$

Let us recall the standard interior trace and boundary traction operators

$$\gamma_0 : \left(H^1(\Omega)\right)^3 \to \left(H^{1/2}(\Gamma)\right)^3 \quad \text{and} \quad \gamma_1 : \left(H^1_{\mathscr{L}}(\Omega)\right)^3 \to \left(H^{-1/2}(\Gamma)\right)^3,$$

respectively, where $H^{1/2}(\Gamma)$ denotes the trace space of $H^1(\Omega)$, $H^{-1/2}(\Gamma)$ is the dual space to $H^{1/2}(\Gamma)$ with respect to the $L^2(\Gamma)$ scalar product, and

$$\left(H^1_{\mathscr{L}}(\Omega)\right)^3 = \left\{\mathbf{v} \in \left(H^1(\Omega)\right)^3 : \ \mathscr{L}\mathbf{v} \in \left(L^2(\Omega)\right)^3\right\}.$$

It is well known (see, e.g., [2–4]) that for any $\mathbf{u} \in \left(H^1_{\mathscr{L}}(\Omega)\right)^3$, there exists the Dirichlet–Neumann map

$$\gamma_1 \mathbf{u} = S\gamma_0 \mathbf{u} - N\mathscr{L}\mathbf{u} \quad \text{on} \ \ \Gamma$$

with the Steklov–Poincaré operator

$$S = (\sigma I + K')V^{-1}(\sigma I + K) + D : \ \left(H^{1/2}(\Gamma)\right)^3 \to \left(H^{-1/2}(\Gamma)\right)^3, \qquad (14.19)$$

where $\sigma(\mathbf{x}) = 1/2$ for almost all $\mathbf{x} \in \Gamma$, and the Newton operator

$$N = V^{-1}N_0 : \ \left(L^2(\Omega)\right)^3 \to \left(H^{-1/2}(\Gamma)\right)^3. \qquad (14.20)$$

In (14.19) and (14.20) we use the single-layer potential operator V, the double-layer potential operator K, the adjoint double-layer potential operator K', and the hypersingular integral operator D given for $\mathbf{x} \in \Gamma$ and $i = 1, 2, 3$ by

$$(V\mathbf{t})_i(\mathbf{x}) = \int_\Gamma \mathbf{t}(\mathbf{y}) \cdot \mathbf{U}_i(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\Gamma_{\mathbf{y}}, \quad V : \ \left(H^{-1/2}(\Gamma)\right)^3 \to \left(H^{1/2}(\Gamma)\right)^3,$$

$$(K\mathbf{u})_i(\mathbf{x}) = \int_\Gamma \mathbf{u}(\mathbf{y}) \cdot \gamma_{1,\mathbf{y}}\mathbf{U}_i(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\Gamma_{\mathbf{y}}, \quad K : \ \left(H^{1/2}(\Gamma)\right)^3 \to \left(H^{1/2}(\Gamma)\right)^3,$$

$$(K'\mathbf{t})_i(\mathbf{x}) = \int_\Gamma \mathbf{t}(\mathbf{y}) \cdot \gamma_{1,\mathbf{x}}\mathbf{U}_i(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\Gamma_{\mathbf{y}}, \quad K' : \ \left(H^{-1/2}(\Gamma)\right)^3 \to \left(H^{-1/2}(\Gamma)\right)^3,$$

$$(D\mathbf{u})_i(\mathbf{x}) = -\gamma_{1,\mathbf{x}} \int_\Gamma \mathbf{u}(y) \cdot \gamma_{1,\mathbf{y}}\mathbf{U}_i(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\Gamma_{\mathbf{y}}, \quad D : \ \left(H^{1/2}(\Gamma)\right)^3 \to \left(H^{-1/2}(\Gamma)\right)^3,$$

and the Newton potential operator N_0 given for $\mathbf{x} \in \Gamma$ and $i = 1, 2, 3$ by

$$(N_0\mathbf{f})_i(\mathbf{x}) = \int_\Omega \mathbf{f}(\mathbf{y}) \cdot \mathbf{U}_i(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\Omega_{\mathbf{y}}, \qquad N_0 : \ \left(L^2(\Omega)\right)^3 \to \left(H^{1/2}(\Gamma)\right)^3,$$

where $\mathbf{U}_i$ are the components of the fundamental solution U of $\mathscr{L}$ (Kelvin tensor),

$$\mathbf{U}_i = [U_{i1}, U_{i2}, U_{i3}]^T, \quad i = 1, 2, 3, \quad \mathsf{U} = \mathsf{U}^T = [\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3],$$

$$U_{ij}(\mathbf{x}, \mathbf{y}) = \frac{1+\nu}{8\pi E(1-\nu)} \left((3-4\nu)\frac{\delta_{ij}}{\|\mathbf{x}-\mathbf{y}\|} + \frac{(x_i - y_i)(x_j - y_j)}{\|\mathbf{x}-\mathbf{y}\|^3} \right), \quad \delta_{ij} = [\mathsf{I}]_{ij}.$$

The mapping properties of the above integral operators are well known [3, 4], in particular, the single-layer potential operator V is $\left(H^{-1/2}(\Gamma)\right)^3$–elliptic, so its inverse exists. We shall need the following lemmas:

Lemma 14.1 *The Steklov–Poincaré operator S is linear, bounded, symmetric, and semi-elliptic on $\left(H^{1/2}(\Gamma)\right)^3$. Moreover, the kernel of S is equal to the space of linearized rigid body motions, i.e.,*

$$\operatorname{Ker} S = \operatorname{Span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -x_2 \\ x_1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -x_3 \\ x_2 \end{bmatrix}, \begin{bmatrix} x_3 \\ 0 \\ -x_1 \end{bmatrix} \right\}. \tag{14.21}$$

Proof See the book by Steinbach [4]. □

Lemma 14.2 *The Newton operator N is linear and bounded on $\left(L^2(\Omega)\right)^3$.*

Proof See the book by Steinbach [4]. □

14.6 Decomposed Contact Problem on Skeleton

Let us consider the variational formulation of the decomposed multibody contact problem to find $\mathbf{u} \in \mathscr{K}_{DD}$ such that

$$q(\mathbf{u}) \le q(\mathbf{v}), \quad \mathbf{v} \in \mathscr{K}_{DD}, \tag{14.22}$$

where

$$q(\mathbf{v}) = \frac{1}{2} a(\mathbf{v}, \mathbf{v}) - \ell(\mathbf{v})$$

and $\mathscr{K}_{DD}$ were introduced in Sect. 11.4. Let us recall that

$$\begin{aligned}
a(\mathbf{u}, \mathbf{v}) &= \sum_{p=1}^{s} a^p(\mathbf{u}^p, \mathbf{v}^p), \qquad a^p(\mathbf{u}^p, \mathbf{v}^p) = \int_{\Omega^p} \sigma\left(\mathbf{v}^p\right) : \varepsilon\left(\mathbf{u}^p\right) \, \mathrm{d}\Omega, \\
\ell(\mathbf{v}) &= \sum_{p=1}^{s} \int_{\Gamma_F^p} \mathbf{f}_\Gamma^p \cdot (\mathbf{v}^p - \mathbf{u}^p) \, \mathrm{d}\Gamma + \sum_{p=1}^{s} \int_{\Omega^p} \mathbf{f}^p \cdot (\mathbf{v}^p - \mathbf{u}^p) \, \mathrm{d}\Omega, \\
V^i &= \left\{ \mathbf{v} \in \left(H^1(\Omega^i)\right)^3 : \mathbf{v} = \mathbf{o} \text{ on } \Gamma_U^i \right\}, \quad i = 1, \ldots, s, \\
V_{DD} &= V^1 \times \cdots \times V^s, \\
\mathscr{K}_{DD} &= \left\{ \mathbf{v} \in V_{DD} : [v_n] \le g \text{ on } \Gamma_C^{ij}, \ (i, j) \in \mathscr{S}; \ \mathbf{v}^i = \mathbf{v}^j \text{ on } \Gamma_G^{ij}, \ i, j = 1, \ldots, s \right\},
\end{aligned}$$

where $\mathscr{S}$ denotes the contact coupling set.

To describe the boundary variational formulation of (14.22), let us introduce

$$V_b^i = \{\mathbf{v} \in (H^{1/2}(\Gamma^1))^3 : \mathbf{v} = \mathbf{o} \text{ on } \Gamma_U^i\}, \quad i = 1, \ldots, s,$$
$$V_{DD}^b = V_b^1 \times \cdots \times V_b^s,$$
$$\mathscr{K}_{DD}^b = \{\mathbf{v} \in V_{DD}^b : [v_n] \le g \text{ on } \Gamma_C^{ij},\ (i,j) \in \mathscr{S};\ \mathbf{v}^i = \mathbf{v}^j \text{ on } \Gamma_G^{ij},\ i,j = 1, \ldots, s\}.$$

For any displacement

$$\mathbf{u} = (\mathbf{u}^1, \ldots, \mathbf{u}^s) \in V_{DD}^b,$$

let us define the energy functional by

$$q^b(\mathbf{u}) = \frac{1}{2} a^b(\mathbf{u}, \mathbf{u}) - \ell^b(\mathbf{u}),$$

where a^b is a bilinear form on V_{DD}^b defined by

$$a^b(\mathbf{u}, \mathbf{v}) = \sum_{p=1}^{s} \langle S^p \mathbf{u}^p, \mathbf{v}^p \rangle$$

and ℓ^b is a linear functional on V_{DD}^b given by

$$\ell^b(\mathbf{v}) = \sum_{p=1}^{s} \left(\langle N^p \mathbf{f}^p, \mathbf{v}^p \rangle + \langle \mathbf{f}_{\Gamma_F^p}^p, \mathbf{v}^p \rangle \right).$$

Recall that

$$S^p : \left(H^{1/2}(\Gamma^p)\right)^3 \to \left(H^{-1/2}(\Gamma^p)\right)^3 \quad \text{and} \quad N^p : \left(L^2(\Omega^p)\right)^3 \to \left(H^{-1/2}(\Gamma^p)\right)^3$$

denote the (local) Steklov–Poincaré and Newton operators, respectively.

The boundary formulation of the decomposed problem (14.22) now reads: find the displacement $\mathbf{u} \in \mathscr{K}_{DD}^b$ such that

$$q^b(\mathbf{u}) \le q^b(\mathbf{u}) \text{ for } \mathbf{v} \in \mathscr{K}_{DD}^b. \tag{14.23}$$

Due to Lemmas 14.1 and 14.2, the bilinear form a^b is bounded, symmetric, and elliptic on V_{DD}^b and the linear functional ℓ^b is bounded on V_{DD}^b. Thus q^b is a bounded convex quadratic functional on V_{DD}^b and problem (14.23) has a unique solution by Theorems 4.4 and 4.5.

14.7 TBETI Discretization of Contact Problem

The boundary element discretization of contact problems defined on the skeleton of the decomposition (14.23) is again very similar to that defined on Ω. The main new feature, as compared with Sect. 14.4, is that the basis functions are defined on 2D boundaries of subdomains Ω^p and are vector valued. After decomposing each Γ^p into n_p^τ triangles τ_i^p, $p = 1, \ldots, s, i = 1, \ldots, n_p^\tau$, we can use, e.g., the constant basis functions $\boldsymbol{\psi}_i^p$ associated with τ_i^p or the linear basis functions $\boldsymbol{\phi}_j^p$ associated with the vertices of τ_i^p (see Fig. 14.2). We denote the number of the basis functions by n_p and look for an approximate solution $\mathbf{u}_h$ in the trial space

$$V_h = V_h^1 \times \cdots \times V_h^s, \quad V_h^p = \text{Span}\left\{\boldsymbol{\phi}_1^p, \ldots, \boldsymbol{\phi}_{n_p}^p\right\}.$$

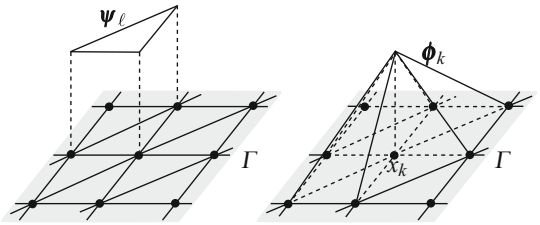

Fig. 14.2 Piecewise constant (*left*) and linear continuous (*right*) basis functions on the surface

After substituting into the forms a^b and ℓ^b, we get the matrix of the local discretized Poincaré–Steklov operators S_p^b and the blocks $\mathbf{f}_p^b$ of the discretized traction

$$\left[\mathsf{S}_p^b\right]_{\ell m} = \langle S^p \boldsymbol{\phi}_\ell^p, \boldsymbol{\phi}_m^p \rangle,$$
$$\left[\mathbf{f}_p^b\right]_\ell = \langle N^p \mathbf{f}^p, \boldsymbol{\phi}_\ell^p \rangle + \langle \mathbf{f}_{\Gamma_F^p}^p, \boldsymbol{\phi}_\ell^p \rangle.$$

To evaluate the entries of the boundary element matrices, we have to approximate (possibly singular) double surface integrals. This can be carried out effectively, e.g., by the semi-analytical approach introduced by Rjasanow and Steinbach in [5], where the inner integral is calculated analytically and the outer one is approximated by using a suitable numerical scheme. See also Steinbach [4, Chap. 10]. If we denote by B_I^b and B_E^b the full rank matrices which describe the discretized non-penetration and gluing conditions, respectively, we get the discretized version of problem (14.23) with auxiliary domain decomposition in the form

$$\min \frac{1}{2}\mathbf{u}^T \mathsf{S}^b \mathbf{u} - (\mathbf{f}^b)^T \mathbf{u} \quad \text{s.t.} \quad \mathsf{B}_I^b \mathbf{u} \le \mathbf{c} \quad \text{and} \quad \mathsf{B}_E^b \mathbf{u} = \mathbf{o}, \tag{14.24}$$

where

$$\mathsf{S}^b = \begin{bmatrix} \mathsf{S}_1^b & \mathsf{O} & \cdots & \mathsf{O} \\ \mathsf{O} & \mathsf{S}_2^b & \cdots & \mathsf{O} \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{O} & \mathsf{O} & \cdots & \mathsf{S}_s^b \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_s \end{bmatrix}, \quad \mathbf{f}^b = \begin{bmatrix} \mathbf{f}_1^b \\ \vdots \\ \mathbf{f}_s^b \end{bmatrix}.$$

The SPS matrix $\mathsf{S}^b \in \mathbb{R}^{n \times n}$ denotes a discrete analog of the Steklov–Poincaré operator that assigns each vector $\mathbf{u}$ of nodal boundary displacements a vector of corresponding nodal forces. Our notation indicates that S^b is closely related to the Schur complement S introduced in Chap. 10. In (14.24), the matrices B_I^b and B_E^b can be obtained from the matrices B_E and B_I from (10.12) by deleting the columns the indices of which correspond to the inner nodes of the subdomains.

Remark 14.1 Since the matrices B_I^b and B_E^b can be obtained from the matrices B_I and B_E by deleting the columns whose indices correspond to the inner nodes of subdomains, it is always possible to achieve that the rows of B are orthonormal provided each node is involved in at most one inequality. This is always possible for two bodies or any number of smooth bodies. To simplify the formulation of the optimality results, we shall assume in what follows (except Chap. 15) that

$$\mathsf{B}^b \left(\mathsf{B}^b\right)^T = \mathsf{I}. \tag{14.25}$$

See also Remark 11.1.

14.8 Dual Formulation

Since the problem (14.24) arising from the application of the TBETI method to the frictionless contact problem has the same structure as that arising from the application of the TFETI method in Chap. 11, we shall reduce our exposition to a brief overview of the basic steps—more details can be found in Chap. 11.

Recall that the Lagrangian associated with the problem (14.24) reads

$$L(\mathbf{u}, \boldsymbol{\lambda}_I, \boldsymbol{\lambda}_E) = \frac{1}{2}\mathbf{u}^T \mathsf{S}^b \mathbf{u} - \mathbf{u}^T \mathbf{f}^b + \boldsymbol{\lambda}_I^T (\mathsf{B}_I^b \mathbf{u} - \mathbf{c}) + \boldsymbol{\lambda}_E^T \mathsf{B}_E^b \mathbf{u},$$

where $\boldsymbol{\lambda}_I$ and $\boldsymbol{\lambda}_E$ are the Lagrange multipliers associated with the inequalities and equalities, respectively. Introducing the notation

$$\boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{\lambda}_I \\ \boldsymbol{\lambda}_E \end{bmatrix}, \quad \mathsf{B}^b = \begin{bmatrix} \mathsf{B}_I^b \\ \mathsf{B}_E^b \end{bmatrix}, \quad \text{and} \quad \mathbf{c} = \begin{bmatrix} \mathbf{c}_I \\ \mathbf{o}_E \end{bmatrix},$$

we can write the Lagrangian briefly as

$$L(\mathbf{u}, \boldsymbol{\lambda}) = \frac{1}{2}\mathbf{u}^T \mathsf{S}^b \mathbf{u} - \mathbf{u}^T \mathbf{f}^b + \boldsymbol{\lambda}^T (\mathsf{B}^b \mathbf{u} - \mathbf{c}).$$

The next steps are the same as in Sect. 11.6. In particular, we get that if $(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}})$ is a KKT couple of the problems (14.18) or (14.24), then $\overline{\boldsymbol{\lambda}}$ solves the minimization problem

$$\min \Theta(\boldsymbol{\lambda}) \quad \text{s.t.} \quad \boldsymbol{\lambda}_I \geq \mathbf{o} \quad \text{and} \quad (\mathsf{R}^b)^T(\mathbf{f}^b - (\mathsf{B}^b)^T\boldsymbol{\lambda}) = \mathbf{o}, \tag{14.26}$$

where

$$\Theta(\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\lambda}^T\mathsf{B}^b(\mathsf{S}^b)^+(\mathsf{B}^b)^T\boldsymbol{\lambda} - \boldsymbol{\lambda}^T(\mathsf{B}^b(\mathsf{S}^b)^+\mathbf{f}^b - \mathbf{c}),$$

and R^b denotes the matrix the columns of which span the kernel of S^b. The matrix R^b can be obtained from the matrix R which spans the kernel of the stiffness matrix K obtained by the volume finite element discretization by deleting the rows which correspond to the nodal variables in the interior of the subdomains.

Let us denote

$$\begin{aligned} &\widetilde{\mathsf{F}}^b = \mathsf{B}^b(\mathsf{S}^b)^+(\mathsf{B}^b)^T, && F = \|\widetilde{\mathsf{F}}^b\|, \\ &\mathsf{F}^b = F^{-1}\widetilde{\mathsf{F}}^b, && \widetilde{\mathbf{d}}^b = F^{-1}\left(\mathsf{B}^b(\mathsf{S}^b)^+\mathbf{f}^b - \mathbf{c}\right), \\ &\widetilde{\mathsf{G}} = (\mathsf{B}^b\mathsf{R}^b)^T, && \widetilde{\mathbf{e}} = (\mathsf{R}^b)^T\mathbf{f}^b, \end{aligned}$$

where T denotes a regular matrix that defines the orthonormalization of the rows of $\widetilde{\mathsf{G}}$ so that the matrix

$$\mathsf{G} = \mathsf{T}\widetilde{\mathsf{G}}$$

has orthonormal rows. After denoting

$$\mathbf{e} = \mathsf{T}\widetilde{\mathbf{e}},$$

problem (14.26) reads

$$\min \quad \frac{1}{2}\boldsymbol{\lambda}^T\mathsf{F}^b\boldsymbol{\lambda} - \boldsymbol{\lambda}^T\widetilde{\mathbf{d}}^b \quad \text{s.t.} \quad \boldsymbol{\lambda}_I \geq \mathbf{o} \quad \text{and} \quad \mathsf{G}\boldsymbol{\lambda} = \mathbf{e}. \tag{14.27}$$

Next we shall transform the problem of minimization on the subset of the affine space to that on the subset of a vector space by means of an arbitrary $\widetilde{\boldsymbol{\lambda}}$ which satisfies

$$\mathsf{G}\widetilde{\boldsymbol{\lambda}} = \mathbf{e}.$$

If the problem (14.27) is a dual problem arising from the discretization of a coercive problem, then we can use Lemma 11.1 to get $\widetilde{\boldsymbol{\lambda}}$ so that

$$\widetilde{\boldsymbol{\lambda}}_I \geq \mathbf{o}.$$

Looking for the solution of (14.27) in the form $\boldsymbol{\lambda} = \boldsymbol{\mu} + \widetilde{\boldsymbol{\lambda}}$, we get, after substituting back $\boldsymbol{\lambda}$ for $\boldsymbol{\mu}$, the problem to find

$$\min \quad \frac{1}{2}\boldsymbol{\lambda}^T \mathsf{F}^b \boldsymbol{\lambda} - \boldsymbol{\lambda}^T \mathbf{d}^b \quad \text{s.t.} \quad \mathsf{G}\boldsymbol{\lambda} = \mathbf{o} \quad \text{and} \quad \boldsymbol{\lambda}_{\mathscr{I}} \geq \ell_I = -\widetilde{\boldsymbol{\lambda}}_I \tag{14.28}$$

with

$$\mathbf{d}^b = \widetilde{\mathbf{d}}^b - \mathsf{F}^b \widetilde{\boldsymbol{\lambda}}.$$

Our final step is based on the observation that problem (14.28) is equivalent to

$$\min \frac{1}{2}\boldsymbol{\lambda}^T \mathsf{H}^b \boldsymbol{\lambda} - \boldsymbol{\lambda}^T \mathsf{P}\mathbf{d}^b \quad \text{s.t.} \quad \mathsf{G}\boldsymbol{\lambda} = \mathbf{o} \quad \text{and} \quad \boldsymbol{\lambda}_{\mathscr{I}} \geq \ell_I, \tag{14.29}$$

where ρ is an arbitrary positive constant and

$$\mathsf{H}^b = \mathsf{P}\mathsf{F}^b\mathsf{P} + \rho\mathsf{Q}, \quad \mathsf{Q} = \mathsf{G}^T\mathsf{G}, \quad \text{and} \quad \mathsf{P} = \mathsf{I} - \mathsf{Q}.$$

Recall that P and Q denote the orthogonal projectors on the kernel of G and on the image space of G^T, respectively. If $\widetilde{\boldsymbol{\lambda}}_I \geq \mathbf{o}$, then $\mathbf{o}$ is a feasible vector for the problem (14.29).

14.9 Bounds on the Spectrum

First observe that $\mathrm{Im}\mathsf{P}$ and $\mathrm{Im}\mathsf{Q}$ are invariant subspaces of H^b and $\mathrm{Im}\mathsf{P}+\mathrm{Im}\mathsf{Q} = \mathbb{R}^m$, as

$$\mathsf{P} + \mathsf{Q} = \mathsf{I}$$

and for any $\boldsymbol{\lambda} \in \mathbb{R}^m$

$$\mathsf{H}^b\mathsf{P}\boldsymbol{\lambda} = (\mathsf{P}\mathsf{F}^b\mathsf{P} + \rho\mathsf{Q})\mathsf{P}\boldsymbol{\lambda} = \mathsf{P}(\mathsf{F}^b\mathsf{P}\boldsymbol{\lambda}) \quad \text{and} \quad \mathsf{H}^b\mathsf{Q}\boldsymbol{\lambda} = (\mathsf{P}\mathsf{F}^b\mathsf{P} + \rho\mathsf{Q})\mathsf{Q}\boldsymbol{\lambda} = \rho\mathsf{Q}\boldsymbol{\lambda}.$$

It follows that

$$\sigma(\mathsf{H}^b|\mathrm{Im}\mathsf{Q}) = \{\rho\},$$

so it remains to find the bounds on

$$\sigma(\mathsf{H}^b|\mathrm{Im}\mathsf{P}) = \sigma(\mathsf{F}^b|\mathrm{Im}\mathsf{P}).$$

The following lemma reduces the problem to the analysis of the local Schur complements.

Lemma 14.3 *Let there be constants* $0 < c < C$ *such that for each* $\boldsymbol{\lambda} \in \mathbb{R}^m$

$$c\|\boldsymbol{\lambda}\|^2 \leq \|(\mathsf{B}^b)^T\boldsymbol{\lambda}\|^2 \leq C\|\boldsymbol{\lambda}\|^2.$$

Then for each $\boldsymbol{\lambda} \in \mathrm{Im}\mathsf{P}$

$$c \min_{i=1,\ldots,s} \overline{\lambda}_{\min}(\mathsf{S}_i^b)\|\boldsymbol{\lambda}\|^2 \leq \boldsymbol{\lambda}^T \mathsf{F}^b \boldsymbol{\lambda} \leq C \max_{i=1,\ldots,s} \|\mathsf{S}_i^b\| \|\boldsymbol{\lambda}\|^2,$$

where S_i^b *denotes the boundary element stiffness matrix associated with* Γ^i.

Proof See the proof of Lemma 10.2. □

Remark 14.2 Lemma 14.3 indicates that the conditioning of H can be improved by the orthonormalization of the rows of the constraint matrix B^b.

We have reduced the problem to bound the spectrum of H to the analysis of the spectrum of the boundary stiffness matrices. The following result, which is due to Langer and Steinbach, is important in the analysis of the optimality of the presented algorithms.

Lemma 14.4 *Let* H *and* h *denote the diameter and the discretization parameter of a quasi-uniform discretization of a domain* $\Omega \subseteq \mathbb{R}^3$ *with shape regular elements. Assume that the elements for the discretization of* Γ *are defined by the traces of those for the discretization of* Ω. *Let* $\mathsf{S}_{H,h}^b$ *and* $\mathsf{S}_{H,h}$ *denote the boundary stiffness matrix and the Schur complement of the stiffness matrix* K *of a subdomain of* Ω *with respect to its interior variables, respectively.*

Then $\mathsf{S}_{H,h}^b$ *and* $\mathsf{S}_{H,h}$ *are spectrally equivalent, i.e., there are constants* c *and* C *independent of* h *and* H *such that for each* $\boldsymbol{\lambda} \in \mathrm{Im}\mathsf{S}$

$$c\boldsymbol{\lambda}^T \mathsf{S}_{H,h} \boldsymbol{\lambda} \leq \boldsymbol{\lambda}^T \mathsf{S}_{H,h}^b \boldsymbol{\lambda} \leq C\boldsymbol{\lambda}^T \mathsf{S}_{H,h} \boldsymbol{\lambda}. \tag{14.30}$$

Proof See Langer and Steinbach [1]. □

The following theorem is now an easy corollary of Lemmas 14.3 and 14.4.

Theorem 14.4 *Let* $\rho > 0$ *and let* $\mathsf{H}_{\rho,H,h}^b$ *denote the Hessian of the cost function of problem (14.29) resulting from the quasi-uniform discretization of problem (14.15) using shape regular boundary elements with the parameters* H *and* h. *Assume that* B^b *satisfies (14.25).*

Then there are constants c *and* C *independent of* h *and* H *such that for each* $\boldsymbol{\lambda} \in \mathbb{R}^n$

$$c\|\boldsymbol{\lambda}\|^2 \leq \boldsymbol{\lambda}^T \mathsf{H}_{\rho,H,h}^b \boldsymbol{\lambda} \leq C \frac{H}{h} \|\boldsymbol{\lambda}\|^2.$$

Proof Substitute (14.30) into Lemma 14.3, use Lemma 11.2, and take into account the fixed regularization term $\rho\mathsf{Q}$. Notice that F is scaled. □

14.10 Optimality

To show that Algorithm 9.2 (SMALBE-M) with the inner loop implemented by Algorithm 8.2 is optimal for the solution of the problem (or a class of problems) (14.29) arising from the varying discretizations of a given frictionless contact problem, let us introduce, as in Sect. 11.10, a new notation that complies with that used in the analysis of the algorithms in Part II.

Let $\rho > 0$ and $C \geq 2$ denote given constants and let

$$\mathscr{T}_C = \{(H, h) \in \mathbb{R}^2 : \ H/h \leq C\}$$

denote the set of indices. For any $t \in \mathscr{T}_C$, let us define the problem

$$\begin{array}{ll} \mathsf{A}_t = \mathsf{P}\mathsf{F}^{\mathsf{b}}\mathsf{P} + \rho\mathsf{Q}, & \mathbf{b}_t = \mathsf{P}\mathbf{d}^b, \\ \mathsf{B}_t = \mathsf{G}, & \ell_I^t = -\widetilde{\boldsymbol{\lambda}}_I, \end{array}$$

where the vectors and matrices of (14.29) arising from varying discretizations of (14.23) with the parameters H and h, $t = (H, h)$. We shall assume that the discretization satisfies the assumptions of Theorem 14.4. Using the procedure described above, we get for each $t \in \mathscr{T}_C$ the problem

$$\text{minimize } \ f_t(\boldsymbol{\lambda}_t) \ \text{ s.t. } \ \mathsf{B}_t\boldsymbol{\lambda}_t = \mathbf{o} \ \text{ and } \ \boldsymbol{\lambda}_t \geq \ell_I^t \tag{14.31}$$

with

$$f_t(\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\lambda}^T\mathsf{A}_t\boldsymbol{\lambda} - \mathbf{b}_t^T\boldsymbol{\lambda}.$$

We shall assume that the discretization satisfies the assumptions of Theorem 14.4 and that

$$\ell_I^t \leq \mathbf{o}.$$

Using $\mathsf{G}\mathsf{G}^T = \mathsf{I}$, we obtain

$$\|\mathsf{B}_t\| \leq 1. \tag{14.32}$$

Moreover, it follows by Theorem 14.4 that for any $C \geq 2$ there are constants $a^C_{\max} > a^C_{\min} > 0$ such that

$$a^C_{\min} \leq \lambda_{\min}(\mathsf{A}_t) \leq \lambda_{\max}(\mathsf{A}_t) \leq a^C_{\max} \tag{14.33}$$

for any $t \in \mathscr{T}_C$. As above, we denote by $\lambda_{\min}(\mathsf{A}_t)$ and $\lambda_{\max}(\mathsf{A}_t)$ the extreme eigenvalues of A_t. Our optimality result for the solution of the class of problems (14.31) arising from the boundary element discretization of the multibody contact problem (11.15) reads as follows.

Theorem 14.5 *Let $C \geq 2$, $\varepsilon > 0$, and $\rho > 0$ denote given constants, let $\{\boldsymbol{\lambda}_t^k\}$ and $\{\boldsymbol{\mu}_t^k\}$ be generated by Algorithm 9.2 (SMALBE-M) for the class of problems (14.31) with*

$$\|\mathbf{b}_t\| \geq \eta_t > 0, \quad 1 > \beta > 0, \quad M_{t,0} > 0, \quad \text{and} \quad \boldsymbol{\mu}_t^0 = \mathbf{o}.$$

Let Step 1 of Algorithm 9.2 be implemented by means of Algorithm 8.2 (MPRGP) with parameters

$$\Gamma > 0 \quad \text{and} \quad \alpha \in (0, 2/a_{\max}^C),$$

so that it generates the iterates

$$\boldsymbol{\lambda}_t^{k,0}, \boldsymbol{\lambda}_t^{k,1}, \ldots, \boldsymbol{\lambda}_t^{k,l} = \boldsymbol{\lambda}_t^k$$

for the solution of (14.31) starting from $\boldsymbol{\lambda}_t^{k,0} = \boldsymbol{\lambda}_t^{k-1}$ with $\boldsymbol{\lambda}_t^{-1} = \mathbf{o}$, where $l = l_{t,k}$ is the first index satisfying

$$\|\mathbf{g}^P(\boldsymbol{\lambda}_t^{k,l}, \boldsymbol{\mu}_t^k, \rho)\| \leq M\|\mathsf{B}_t\boldsymbol{\lambda}_t^{k,l}\|$$

or

$$\|\mathbf{g}^P(\boldsymbol{\lambda}_t^{k,l}, \boldsymbol{\mu}_t^k, \rho)\| \leq \varepsilon M\|\mathbf{b}_t\|.$$

Then for any $t \in \mathscr{T}_C$ and problem (14.31), an approximate solution $\boldsymbol{\lambda}_t^{k_t}$ which satisfies

$$\|\mathbf{g}^P(\boldsymbol{\lambda}_t^{k_t}, \boldsymbol{\mu}_t^{k_t}, \rho)\| \leq \varepsilon M\|\mathbf{b}_t\| \quad \text{and} \quad \|\mathsf{B}_t\boldsymbol{\lambda}_t^{k_t}\| \leq \varepsilon\|\mathbf{b}_t\|$$

is generated at $O(1)$ matrix–vector multiplications by the Hessian of f_t.

Proof The class of problems satisfies all assumptions of Theorem 9.4 (i.e., $\mathbf{o}$ is feasible and the inequalities (14.32) and (14.33) hold true) for the set of indices $\mathscr{T}_C$. Thus to complete the proof, it is enough to apply Theorem 9.4. □

Since the cost of a matrix–vector multiplication by the Hessian A^t is proportional to the number of dual variables, Theorem 14.5 proves the numerical scalability of TBETI. The (week) parallel scalability is supported by the structure of A_t.

14.11 Numerical Experiments

The algorithms presented here were implemented in `MatSol` [10] and tested on a number of academic benchmarks and real world problems, see, e.g., Sadowská et al. [11] or [12]. Here, we give some results that illustrate the numerical scalability and effectiveness of TBETI using the benchmark of Chap. 12. We also compare the precision of TFETI and TBETI on the solution of a variant of the Hertz problem with the known analytic solution. All the computations were carried out with the

parameters recommended in the description of the algorithms in Chaps. 7–9. The relative precision of the computations was the same as in Chaps. 11 and 12, i.e., 10^{-4} for the academic benchmark and the 10^{-6} for the real world problems.

14.11.1 Academic Benchmark

We consider the contact problems of two cantilever beams in contact with friction that was introduced in Sect. 12.9. The problems were decomposed and discretized with varying decomposition and discretization parameters h and H as in Sect. 12.9. We kept $H/h = 12$, so that the assumptions of Theorem 14.5 are satisfied. The performance of the algorithms is documented in the following graphs.

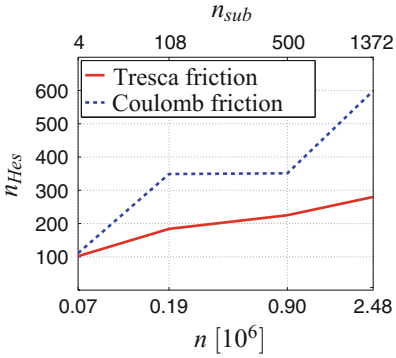

Fig. 14.3 Tresca and Coulomb BETI: numerical scalability of cantilever beams—matrix–vector multiplications by F

The numbers of outer iterations and the multiplications by the Hessian F of the dual function depending on the primal dimension n and the number of subdomains n_{sub} are in Fig. 14.3, both for the problem with the Tresca and Coulomb friction with the friction coefficient $\Phi = 0.1$. We can see a stable number of outer iterations and mildly increasing number of inner iterations for n ranging from 7224 to 2,458,765. The dual dimension of the problems ranged from 1859 to 1,261,493. We conclude that the performance of the algorithms based on TBETI is very similar to those based on TFETI and is in agreement with the theory. Notice that the primal dimension of the problem discretized by BETI is typically much smaller than that discretized by BETI with the same discretization parameter h.

14.11.2 Comparison TFETI and TBETI

We consider a frictionless 3D Hertz problem depicted in Fig. 14.4, with the Young modulus $2.1 \cdot 10^5$ MPa and the Poisson ratio 0.3. The ball is loaded on the top by the force $F = -5000$ [N]. The ANSYS discretization of the two bodies was decomposed by METIS into 1024 subdomains.

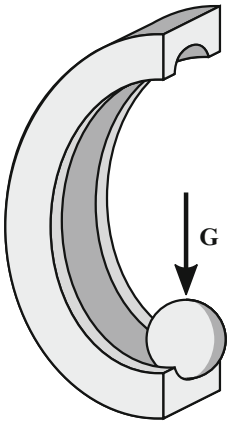

Fig. 14.4 Model specification

The comparison of TFETI and TBETI in terms of computational times and the number of Hessian multiplications is given in Table 14.1.

Table 14.1 Numerical performance of TFETI and TBETI applied to the Hertz problem

Method	Number of primal DOFs	Number of dual DOFs	Preprocessing time	Solution time	Number of Hessian applications
TFETI	4,088,832	926,435	21 min	1 h 49 min	593
TBETI	1,849,344	926,435	1 h 33 min	1 h 30 min	667

We can see that the time of computation required by the two methods are comparable, especially if we take into account that the preprocessing time can be reduced very efficiently on more processors than 24 used in the test.

In Fig. 14.5 we can see a fine correspondence of the contact pressures computed by TFETI and TBETI with the analytical solution. We can see that BETI returns a bit more precise results, which is not surprising, as TFETI uses the exact solutions in the interior of subdomains. The convergence criterion was the reduction of the norm of the projected gradient of the dual energy function by 10^{-6}.

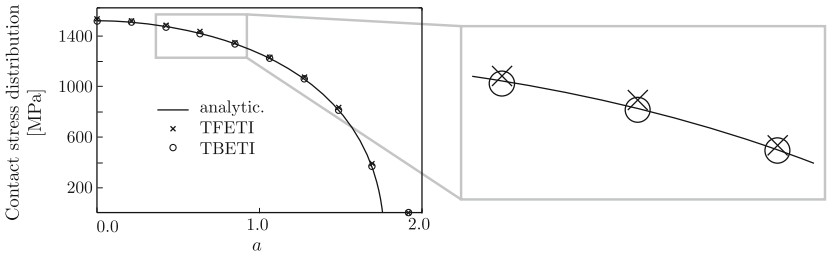

Fig. 14.5 Correspondence of numerical Hertz contact pressures to the analytic solution

14.11.3 Ball Bearing

We solved also the contact problem of ball bearing depicted in Figs. 1.1 and 1.3. We imposed the Dirichlet boundary conditions on the outer ring and loaded the opposite part of the inner diameter with the force 4500 [N]. The discretized geometry was decomposed into 960 subdomains by Metis. Numerical comparison of TFETI and TBETI is in Table 14.2.

Table 14.2 Numerical performance of TFETI and TBETI applied to the ball bearing problem

Method	Number of primal DOFs	Number of dual DOFs	Preprocessing time	Solution time	Number of Hessian applications
TFETI	1,759,782	493,018	129 s	2 h 5 min	3203
TBETI	1,071,759	493,018	715 s	1 h 5 min	2757

For more information see also [11, 12].

14.12 Comments

For a nice introductory course on the classical boundary element method which covers also some more advanced topics, see Gaul, Kögl, and Wagner [7]. For more formal exposition, see the books by Steinbach [4] or McLean [2]. Engineering applications can be found in the book edited by Langer et al. [13].

The introduction of BETI method by Langer and Steinbach [1] opened the way to the development of scalable domain decomposition-based algorithms for the solution of problems discretized by the boundary element method. A variant of BETI which we use here appeared first in Thesis by Of [14] as AF (all floating) BETI. See also Of and Steinbach [15]. Here we call it TBETI (Total) as it is shorter and indicates a close relation to TFETI, which was developed independently at about the same time.

The scalability results reported here were published first for a scalar model problem in Bouchala et al. [8], then for 3D multibody elastic problems in Bouchala et al. [8] and Sadowská et al. [11]. The scalability of the algorithm adapted for the solution of problems with Tresca (given) friction is reported in Sadowská et al. [12]. We are not aware of other scalable methods for the solution of contact problems by the boundary element method.

The performance of any algorithm which uses BEM can be improved by so-called fast methods (see, e.g., Rjasanow and Steinbach [5]). They include the hierarchical matrices introduced by Hackbusch [16] (see a comprehensive exposition in Bebendorf [17]), the Adaptive Cross Approximation (ACA) (see, e.g., Bebendorf [18] or Bebendorf and Rjasanow [17]), or Fast Multipole Method (FMM) (see, e.g., Greengard and Rokhlin [19, 20] or Of, Steinbach, and Rokhlin [21]). These methods accelerate the evaluation of the matrices and the consequent matrix–vector multiplication and lead to asymptotically nearly linear space and time complexities.

References

1. Langer, U., Steinbach, O.: Boundary element tearing and interconnecting methods. Computing **71**, 205–228 (2003)
2. McLean, W.: Strongly Elliptic Systems and Boundary Integral Equations. Cambridge University Press, Cambridge (2000)
3. Costabel, M.: Boundary integral operators on Lipschitz domains: elementary results. SIAM J. Math. Anal. **19**, 613–626 (1988)
4. Steinbach, O.: Numerical Approximation Methods for Elliptic Boundary Value Problems. Finite and Boundary Elements. Springer, New York (2008)
5. Rjasanow, S., Steinbach, O.: The Fast Solution of Boundary Integral Equations. Mathematical and Analytical Techniques with Applications to Engineering. Springer, New York (2007)
6. Steinbach, O.: Stability Estimates for Hybrid Coupled Domain Decomposition Methods. Lecture Notes in Mathematics. Springer, Berlin (2003)
7. Gaul, L., Kögl, M., Wagner, M.: Boundary Elements for Engineers and Scientists. Springer, Berlin (2003)
8. Bouchala, J., Dostál, Z., Sadowská, M.: Theoretically supported scalable BETI method for variational inequalities. Computing **82**, 53–75 (2008)
9. Dostál, Z., Friedlander, A., Santos, S.A., Malík, J.: Analysis of semicoercive contact problems using symmetric BEM and augmented Lagrangians. Eng. Anal. Bound. Elem. **18**(3), 195–201 (1996)
10. Kozubek, T., Markopoulos, A., Brzobohatý, T., Kučera, R., Vondrák, V., Dostál, Z.: MatSol–MATLAB efficient solvers for problems in engineering. http://industry.it4i.cz/en/products/matsol/ (2015)
11. Sadowská, M., Dostál, Z., Kozubek, T., Markopoulos, A., Bouchala, J.: Scalable Total BETI based solver for 3D multibody frictionless contact problems in mechanical engineering. Eng. Anal. Bound. Elem. **35**, 330–341 (2011)
12. Sadowská, M., Dostál, Z., Kozubek, T., Markopoulos, A., Bouchala, J.: Engineering multibody contact problems solved by scalable TBETI. In: Fast Boundary Element Methods in Engineering and Industrial Applications, LNACM, vol. 63, pp. 241–269. Springer, Berlin (2012)
13. Langer, U., Schanz, M., Steinbach, O., Wendland, W.L. (eds.): Fast Boundary Element Methods in Engineering and Industrial Applications. Lecture Notes in Applied and Computational MechanicsSeries, vol. 63. Springer, Berlin (2012)

14. Of, G.: BETI - Gebietszerlegungsmethoden mit schnellen Randelementverfahren und Anwendungen. Ph.D. Thesis, University of Stuttgart (2006)
15. Of, G., Steinbach, O.: The all-floating boundary element tearing and interconnecting method. J. Numer. Math. **17**(4), 277–298 (2009)
16. Hackbusch, W.: A sparse matrix arithmetic based on H-matrices. Part I: introduction to H-matrices. Computing **62**, 89–108 (1999)
17. Bebendorf, M.: Hierarchical Matrices: A Means to Efficiently Solve Elliptic Boundary Value Problems. Lecture Notes in Computational Science and Engineering, vol. 63. Springer, Berlin (2008)
18. Bebendorf, M., Rjasanow, S.: Adaptive low-rank approximation of collocation matrices. Computing **70**, 1–24 (2003)
19. Greengard, L.F., Rokhlin, V.: A fast algorithm for particle simulations. J. Comput. Phys. **73**, 325–348 (1987)
20. Greengard, L.F., Rokhlin, V.: A new version of the fast multipole method for the Laplace equation in three dimensions. Acta Numerica, Cambridge University Press, 229–269 (1997)
21. Of, G., Steinbach, O., Wendland, W.L.: Applications of a fast multipole Galerkin in boundary element method in linear elastostatics. Comput. Vis. Sci. **8**, 201–209 (2005)

Chapter 15
Mortars

The theoretical results on numerical scalability of algorithms for the solution of contact problems presented in Chaps. 10–14 relied essentially on the strong linear independence of the rows of constraint matrices that describe "gluing" of subdomains and non-penetration conditions. The constraint matrices with the desired properties assumed matching grids and node-to-node linearized non-penetration conditions, possibly obtained by an effective orthogonalization of the equality constraints. However, such approach has poor approximation properties and causes problems when the contact conditions are imposed on large curved interfaces or non-matching grids—the latter are inevitably present in the solution of transient contact problems.

Similar to the description of conditions of equilibrium, most of the drawbacks mentioned above can be overcome when we impose the non-penetration conditions in average. Our basic tool will be so called mortars, which enforce the non-penetration by the discretization of week non-penetration conditions like (13.14), in particular the biorthogonal mortars introduced by Wohlmuth [1]. The latter are supported by a nice approximation theory of variationally consistent discretization (see Wohlmuth [2]) and can be effectively implemented.

Until recently, it was not clear whether the mortars can be plugged into the FETI-based domain decomposition algorithms in a way which preserves the scalability of the algorithms. The reasons were that the constraint matrices resulting from the mortar discretization do not have a nice block diagonal structure like those arising from the node-to-node schemes and that it was not clear how to orthogonalize effectively the rows corresponding to the inequality constraints against those enforcing the "gluing" of the subdomains, especially those associated with the "wire baskets."

In this chapter, we briefly review the description of non-penetration conditions by means of biorthogonal mortars, prove that the constraint matrices arising from the discretization by some biorthogonal bases are well conditioned under natural restrictions, and show that the procedure complies well with the FETI domain decomposition method. The theoretical results are illustrated by numerical experiments.

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3_15

15.1 Variational Non-penetration Conditions

As in Sect. 11.1, let us consider a system of bodies which occupy in the reference configuration open bounded domains $\Omega^1, \ldots, \Omega^s \subset \mathbb{R}^3$ with the Lipchitz boundaries $\Gamma^1, \ldots, \Gamma^s$ and their parts Γ_F^i, Γ_U^i, and Γ_C^i. Suppose that some nonempty Γ_C^p comprises a part $\Gamma_C^{pq} \subseteq \Gamma_C^p$ that can get into contact with $\overline{\Omega}^q$ as in Fig. 11.1. We assume that Γ_C^p is sufficiently smooth, so that there is a well-defined outer normal $\mathbf{n}^p(\mathbf{x})$ at almost each $\mathbf{x} \in \Gamma_C^p$. From each couple of indices $\{p, q\}$ which identify $\Gamma_C^{pq} \neq \emptyset$, we choose one index to identify the slave side of a possible contact interface and define a contact coupling set $\mathscr{S}$. Thus if $(p, q) \in \mathscr{S}$, then $\Gamma_C^{pq} \neq \emptyset$ is the slave side of the contact interface that can come into contact with the master side Γ_C^{qp}.

We shall first discuss the problems that can be explained without loss of generality on two bodies assuming that $\mathscr{S} = \{(1, 2)\}$, so that we can simplify the notation to $\Gamma_C^1 = \Gamma_C^{12}$ and $\Gamma_C^2 = \Gamma_C^{21}$. We shall use the notation introduced in Sect. 11.2 to write the contact conditions briefly as

$$[u_n] \leq g, \quad \lambda_n \geq 0, \quad \lambda_n([u_n] - g) = 0, \quad \boldsymbol{\lambda} = \lambda_n \mathbf{n}, \quad \mathbf{x} \in \Gamma_C^1, \tag{15.1}$$

where $\mathbf{n} = \mathbf{n}(\mathbf{x})$ denotes an outer normal unit vector at $\mathbf{x} \in \Gamma_C^1$, $\boldsymbol{\lambda}$ is the contact traction,

$$[u_n] = (\mathbf{u}^1 - \mathbf{u}^2 \circ \chi) \cdot \mathbf{n}^1$$

denotes the jump in normal displacements, $\chi = \chi^{12} : \Gamma_C^1 \mapsto \Gamma_C^2$ denotes the one-to-one slave–master mapping and

$$g = g(\mathbf{x}) = \big(\chi(\mathbf{x}) - \mathbf{x}\big) \cdot \mathbf{n}(\mathbf{x})$$

is the initial normal gap. The first of the conditions (15.1) defines the nonempty, closed, and convex subset $\mathscr{K}$ of the Hilbert space

$$V = V^1 \times V^2, \quad V^1 = (H^1(\Omega^1))^3, \quad V^2 = (H^1(\Omega^2))^3$$

by

$$\mathscr{K} = \{\mathbf{v} \in V : [v_n] \leq g \quad \text{on} \quad \Gamma_C^1\}. \tag{15.2}$$

The feasible set can be alternatively characterized by means of the dual cone. To describe the weak non-penetration conditions, let us denote

$$W = H^{1/2}\big(\Gamma_C^1\big), \quad W^+ = \{w \in W : w \geq 0\},$$
$$M = H^{-1/2}(\Gamma_C^1), \quad M^+ = \{\mu \in M : \langle \mu, w \rangle \geq 0, \ w \in W^+\},$$

so that M denotes the dual space of the trace space W and M^+ denotes the dual cone. The feasible set (with respect to linearized non-penetration conditions) can be characterized by

$$\mathcal{K} = \{\mathbf{v} \in V \ : \ \langle \mu, [v_n] \rangle_{\Gamma_C^1} \leq \langle \mu, g \rangle_{\Gamma_C^1} \ \text{ for all } \mu \in M^+\}.$$

15.2 Variationally Consistent Discretization

Let us assume for simplicity that the domains Ω^1 and Ω^2 are polyhedral, so that there are independent families of shape regular triangulations $\mathcal{T}^k$, $k \in \{1, 2\}$, such that $\overline{\Omega}^k = \bigcup_{\omega \in \mathcal{T}^k} \overline{\omega}$. Let $\mathcal{F}^k$ denote the set of contact faces τ of the elements of $\mathcal{T}^k$, so that $\mathcal{F}^k$ defines a 2-dimensional surface mesh of Γ_C^k, $\overline{\Gamma}_C^k = \bigcup_{\tau \in \mathcal{F}^k} \overline{\tau}$, $k \in \{1, 2\}$. The surface mesh on Γ_C^2 is mapped by χ^{-1} onto Γ_C^1, resulting in possibly non-matching meshes on the contact interface.

Following Wohlmuth [2], we use the standard low order conforming finite elements for the displacements and the dual finite elements which reproduce constants for surface traction. We shall use the notation

$$V_h = V_h^1 \times V_h^2, \qquad V_h^k = \text{Span}\{\phi_p : p \in \mathcal{P}^k\}^3, \ \ k \in \{1, 2\}, \tag{15.3}$$

$$M_h^+ = \left\{ \mu = \sum_{p \in \mathcal{P}_C^1} \beta_p \psi_p, \ \beta_p \in \mathbb{R}^+ \right\}, \tag{15.4}$$

where $\mathcal{P}^k$ is the set of all vertices of $\mathcal{T}^k$, $\mathcal{P}_C^k$ is the set of all vertices which belong to $\overline{\Gamma}_C^k$, $k \in \{1, 2\}$, ϕ_p denotes the standard conforming nodal basis function associated with the vertex p, and ψ_p denotes the dual basis function associated with p. In particular, for any $\mathbf{x} \in \Gamma_C^k$

$$\sum_{p \in \mathcal{P}^k} \phi_p(\mathbf{x}) = 1, \ \ k = 1, 2. \tag{15.5}$$

The following properties of the dual basis functions ψ_p are essential for our presentation.

- The support of ψ_p is local, i.e.,

$$\text{supp } \psi_p = \text{supp } \phi_p | \Gamma_C^1, \quad p \in \mathcal{P}_C^1. \tag{15.6}$$

- The basis functions ψ_p and ϕ_p are locally biorthogonal, i.e.,

$$\int_\tau \phi_p \psi_q \, \mathrm{d}\Gamma = \delta_{pq} \int_\tau \phi_p \mathrm{d}\Gamma, \quad p, q \in \mathcal{P}_C^1, \ \tau \in \mathcal{F}^1, \tag{15.7}$$

where $\mathscr{F}^1$ denotes the set of all contact faces on the slave side. The basis functions are required to enjoy two other properties, *the best approximation property* and *the uniform inf–sup condition* that are essential for the development of the approximation theory [1]. Notice that there exists no set of nonnegative basis functions that satisfy (15.7), so M_h^+ is not a subset of M^+. Observing that each $\mathbf{v} \in V_h$ can be written in the form

$$\mathbf{v} = \sum_{p \in \mathscr{P}^1} \phi_p \mathbf{x}_p + \sum_{p \in \mathscr{P}^2} \phi_p \mathbf{x}_p, \quad \mathbf{x}_p \in \mathbb{R}^3,$$

we can write the non-penetration condition in more detail as

$$\int_{\Gamma_C^1} \psi_p \left(\sum_{q \in \mathscr{P}_C^1} \phi_q \mathbf{x}_q - \sum_{q \in \mathscr{P}_C^2} (\phi_q \circ \chi) \mathbf{x}_q \right) \cdot \mathbf{n} \, \mathrm{d}\Gamma \le \int_{\Gamma_C^1} \psi_p g \, \mathrm{d}s, \quad p \in \mathscr{P}_C^1. \tag{15.8}$$

Our approximation of $\mathscr{K}$ reads

$$\mathscr{K}_h = \{\mathbf{v} \in V_h \, : \, \langle \mu, [v_n] \rangle_{\Gamma_C^1} \le \langle \mu, g \rangle_{\Gamma_C^1} \text{ for all } \mu \in M_h^+\}.$$

To write (15.8) in a convenient matrix form, let us assign local indices $1, \ldots, n_1$ and $n_1 + 1, \ldots, n_1 + n_2$ to the vertices on the contact interface on the slave and master side, respectively, and assume that $\mathbf{n}(\mathbf{x}_i) = \mathbf{n}(\mathbf{x})$ for $i = 1, \ldots, n_1$ and $\mathbf{x} \in \operatorname{supp} \phi_i$, so that (15.8) is equivalent to

$$\mathsf{B}_N \mathbf{x} \le \Sigma^{-1} \mathbf{g}_N, \quad \mathsf{B}_N = \Sigma^{-1} \left[\mathsf{D}, -\mathsf{M} \right], \quad \mathbf{x} = [(\mathbf{x}^1)^T, (\mathbf{x}^2)^T]^T,$$

where

$$\mathsf{D} = \begin{bmatrix} d_1 \mathbf{n}_1^T & \ldots & \mathbf{o}^T \\ . & \ldots & . \\ \mathbf{o}^T & \ldots & d_{n_1} \mathbf{n}_{n_1}^T \end{bmatrix}, \qquad d_i = \int_{\Gamma_C^1} \phi_i^1 \, \mathrm{d}s, \tag{15.9}$$

$$\mathsf{M} = \begin{bmatrix} m_{11} \mathbf{n}_1^T & \ldots & m_{1n_2} \mathbf{n}_1^T \\ . & \ldots & . \\ m_{n_1 1} \mathbf{n}_{n_1}^T & \ldots & m_{n_1 n_2} \mathbf{n}_{n_1}^T \end{bmatrix}, \quad m_{ij} = \int_{\Gamma_C^1} \psi_i^1 (\phi_j^2 \circ \chi) \, \mathrm{d}s, \tag{15.10}$$

$$\mathbf{g}_N = [g_i], \qquad g_i = \int_{\Gamma_C^1} \psi_p^1 \, g \, \mathrm{d}s, \quad i = 1, \ldots, n_1, \tag{15.11}$$

and Σ denotes the scaling matrix which normalizes the rows of $[\mathsf{D}, -\mathsf{M}]$, so that the diagonal entries of $\mathsf{B}_N \mathsf{B}_N^T$ are equal to one, i.e.,

$$[\Sigma]_{ii}^2 = d_i^2 + \sum_{j=1}^{n_2} m_{ij}^2, \quad i = 1, \ldots, n_1. \tag{15.12}$$

To simplify the reading, we added a superscript k, $k \in \{1, 2\}$ to the basis functions associated with Γ_C^k. Due to the block structure of B_N, we have

$$\mathsf{B}_N \mathsf{B}_N^T = \Sigma^{-2}\mathsf{D}^2 + \Sigma^{-1}\mathsf{M}\mathsf{M}^T\Sigma^{-1}, \tag{15.13}$$

where D is diagonal SPD and $\mathsf{M}\mathsf{M}^T$ is SPS. Very useful discussion of implementation details and catches can be found in Popp et al. [3].

15.3 Conditioning of Mortar Non-penetration Matrix

Now we are ready to give a bound on the squares of the singular values of B, in particular on the smallest eigenvalue $\lambda_{\min}(\mathsf{B}_N \mathsf{B}_N^T)$ and on the largest eigenvalue $\lambda_{\max}(\mathsf{B}_N \mathsf{B}_N^T)$. To simplify the description of our results, let us introduce the following definition.

Definition 15.1 The dual basis functions ψ_i, ψ_j, $i \neq j$, defined on Γ_C^1 are *near* if there is a finite element basis function ϕ_ℓ^2 defined on Γ_C^2 such that

$$\operatorname{supp}(\phi_\ell^2 \circ \chi) \cap \operatorname{supp} \psi_i \neq \emptyset \quad \text{and} \quad \operatorname{supp}(\phi_\ell^2 \circ \chi) \cap \operatorname{supp} \psi_j \neq \emptyset.$$

The *cover number of a dual basis function* ψ_i is the number of dual basis functions that are near to ψ_i. The *cover number* $\overline{k}$ *of the mortar discretization* is the maximal cover number of the dual basis functions on Γ_C^1.

Theorem 15.1 *Let* B_N *denote the matrix arising from the consistent mortar discretization of the conditions of non-penetration for our model problem with the finite elements that satisfy (15.5). Then*

$$\min_{i=1,\dots,n_1} \frac{\left(\int_{\Gamma_C^1} \psi_i \, \mathrm{d}s\right)^2}{\left(\int_{\Gamma_C^1} \psi_i \, \mathrm{d}s\right)^2 + \left(\int_{\Gamma_C^1} |\psi_i| \, \mathrm{d}s\right)^2} \leq \lambda_{\min}(\mathsf{B}_N \mathsf{B}_N^T) \leq 1 \tag{15.14}$$

and

$$\lambda_{\max}(\mathsf{B}_N \mathsf{B}_N^T) \leq 1 + \overline{k} \max_{i=1,\dots,n_1} \frac{\left(\int_{\Gamma_C^1} |\psi_i| \, \mathrm{d}s\right)^2}{\left(\int_{\Gamma_C^1} \psi_i \, \mathrm{d}s\right)^2 + \left(\int_{\Gamma_C^1} |\psi_i| \, \mathrm{d}s\right)^2}, \tag{15.15}$$

where $\overline{k}$ *denotes the cover number of the discretization.*

Proof First observe that

$$\lambda_{\min}\left(\mathsf{B}_N \mathsf{B}_N^T\right) = \lambda_{\min}\left(\Sigma^{-2}\mathsf{D}^2 + \Sigma^{-1}\mathsf{M}\mathsf{M}^T\Sigma^{-1}\right) \geq \lambda_{\min}\left(\Sigma^{-2}\mathsf{D}^2\right)$$

and that $\Sigma^{-2}\mathsf{D}^2$ is diagonal. Using (15.12) and (15.13), we get

$$\begin{aligned}\left[\Sigma^{-2}\mathsf{D}^2\right]_{ii} &= \frac{d_i^2}{d_i^2 + \sum_{j=1}^{n_2} m_{ij}^2} \\ &= \frac{\left(\int_{\Gamma_C^1} \psi_i \,\mathrm{d}s\right)^2}{\left(\int_{\Gamma_C^1} \psi_i \,\mathrm{d}s\right)^2 + \sum_{j=1}^{n_2} \left(\int_{\Gamma_C^1} \psi_i (\phi_j^2 \circ \chi) \,\mathrm{d}s\right)^2}.\end{aligned}$$

Noticing that the basis functions ϕ_j satisfy (15.5), we can estimate the last sum in the denominator of the last term by

$$\begin{aligned}\sum_{j=1}^{n_2} \left(\int_{\Gamma_C^1} \psi_i (\phi_j^2 \circ \chi) \,\mathrm{d}s\right)^2 &\le \left(\sum_{j=1}^{n_2} \int_{\Gamma_C^1} \left|\psi_i (\phi_j^2 \circ \chi)\right| \mathrm{d}s\right)^2 \\ &= \left(\int_{\Gamma_C^1} \sum_{j=1}^{n_2} |\psi_i| \phi_j^2 \circ \chi \,\mathrm{d}s\right)^2 \\ &= \left(\int_{\Gamma_C^1} |\psi_i| \,\mathrm{d}s\right)^2, \quad i = 1, \ldots, n_1,\end{aligned}$$

which proves the left inequality of (15.14).

To prove the right inequality, just observe that for the vector $\mathbf{e}_i$ with the entries δ_{ij}, we have due to the normalization

$$[\mathsf{B}\mathsf{B}^T]_{ii} = \mathbf{e}_i^T \mathsf{B}_N \mathsf{B}_N^T \mathbf{e}_i = 1.$$

To prove the upper bound (15.15), we recall that the ℓ_∞-norm of any square matrix dominates its Euclidean norm, i.e.,

$$\lambda_{\max}\left(\mathsf{B}_N \mathsf{B}_N^T\right) \le \max_{i=1,\ldots,n_1} \sum_{j=1}^{n_1} |\mathbf{b}_i^T \mathbf{b}_j| = 1 + \max_{i=1,\ldots,n_1} \sum_{j=1, j\neq i}^{n_1} |\mathbf{b}_i^T \mathbf{b}_j|. \tag{15.16}$$

Our next goal will be the upper bound on $|\mathbf{b}_i^T \mathbf{b}_j|$. Using the Cauchy interlacing inequalities (2.21), we get that the eigenvalues of any submatrix

$$\mathsf{T}_{ij} = \begin{bmatrix} 1 & \mathbf{b}_i^T \mathbf{b}_j \\ \mathbf{b}_j^T \mathbf{b}_i & 1 \end{bmatrix}, \qquad i, j \in \{1, \ldots, n_1\},$$

of $\mathsf{B}_N \mathsf{B}_N^T$ satisfy

$$\lambda_{\min}\left(\mathsf{T}_{ij}\right) \ge \lambda_{\min}(\mathsf{B}_N \mathsf{B}_N^T).$$

Direct computations reveal the spectrum $\sigma(\mathsf{T}_{ij})$ of T_{ij}; we get

$$\sigma(\mathsf{T}_{ij}) = \{1 + |\mathbf{b}_i^T \mathbf{b}_j|, 1 - |\mathbf{b}_i^T \mathbf{b}_j|\}.$$

Combining these observations with (15.14), we get

$$\begin{aligned} 1 - |\mathbf{b}_i^T \mathbf{b}_j| \geq \lambda_{\min}(\mathsf{B}_N \mathsf{B}_N^T) &\geq \frac{\left(\int_{\Gamma_C^1} \psi_i \, \mathrm{d}s\right)^2}{\left(\int_{\Gamma_C^1} \psi_i \, \mathrm{d}s\right)^2 + \left(\int_{\Gamma_C^1} |\psi_i| \, \mathrm{d}s\right)^2} \\ &= 1 - \frac{\left(\int_{\Gamma_C^1} |\psi_i| \, \mathrm{d}s\right)^2}{\left(\int_{\Gamma_C^1} \psi_i \, \mathrm{d}s\right)^2 + \left(\int_{\Gamma_C^1} |\psi_i| \, \mathrm{d}s\right)^2}, \end{aligned}$$

so that

$$|\mathbf{b}_i^T \mathbf{b}_j| \leq \frac{\left(\int_{\Gamma_C^1} |\psi_i| \, \mathrm{d}s\right)^2}{\left(\int_{\Gamma_C^1} \psi_i \, \mathrm{d}s\right)^2 + \left(\int_{\Gamma_C^1} |\psi_i| \, \mathrm{d}s\right)^2}. \tag{15.17}$$

To finish the proof, substitute (15.17) to (15.16) and observe that

$$\mathbf{b}_i^T \mathbf{b}_j = \sum_{k=1}^{n_2} \left(\int_{\Gamma_C^1} \psi_i (\phi_k^2 \circ \chi) \, \mathrm{d}s \right) \left(\int_{\Gamma_C^1} \psi_j (\phi_k^2 \circ \chi) \, \mathrm{d}s \right).$$

It follows that if the dual functions ψ_i and ψ_j are not near and $i \neq j$, then $\mathbf{b}_i^T \mathbf{b}_j \neq 0$. □

It is possible to evaluate directly the bounds for the linear elements, so we can formulate the following corollary.

Corollary 15.1 *Let B_N denote the matrix arising from the consistent mortar discretization of the conditions of non-penetration of two 3D bodies using the linear finite element basis functions reproducing constants. Then*

$$\frac{1}{7} < \frac{64}{425} \leq \lambda_{\min}\left(\mathsf{B}_N \mathsf{B}_N^T\right) \leq \|\mathsf{B}_N\|^2 \leq 1 + \overline{k}. \tag{15.18}$$

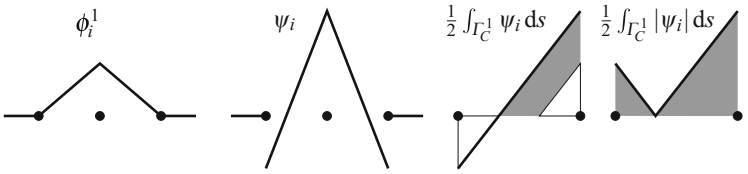

Fig. 15.1 Illustration to Corollary 15.1

Proof The proof of the corollary is based on the observation that the relevant characteristics of the biorthogonal basis functions associated with linear elements do not depend on the geometry. The slave continuous basis functions ϕ_i^1 and the discontinuous dual basis functions ψ_i^1 for 2D problems (see Fig. 15.1) have end–point values $\{1, 0\}$ and $\{2, -1\}$, respectively. To evaluate the estimate (15.14), we need explicitly the ratio between the integrals of ψ_i^1 and $|\psi_i^1|$ over Γ_C^1, which can be shown to be 3/5 (see the gray areas in the right part of Fig. 15.1). The corresponding value for the 3D problems is 8/19. Thus

$$\frac{\left(\int_{\Gamma_C^1} \psi_i \, \mathrm{d}s\right)^2}{\left(\int_{\Gamma_C^1} \psi_i \, \mathrm{d}s\right)^2 + \left(\int_{\Gamma_C^1} |\psi_i| \, \mathrm{d}s\right)^2} \geq \frac{1}{1 + \left(\frac{19}{8}\right)^2} = \frac{64}{425}.$$

□

Though the estimate (15.18) captures the qualitative features of the conditioning of mortar constraint matrices, it is rather pessimistic. The following Table 15.1 provides information about the singular values of B_I for non-matching regular discretizations of the contact interface of Hertz's problem with varying discretization parameters. We can see nice numbers for small values of $h_{\text{slave}}/h_{\text{master}}$, which are relevant from the point of view of the approximation theory. See also Vlach, Dostál, and Kozubek [4].

Table 15.1 Conditioning of the mortar matrices B_I for 3D Hertz problem with non-matching grids

$h_{\text{slave}}/h_{\text{master}}$	1/6	1/3	2/3	1	3/2	3	6
$\lambda_{\min}(\mathsf{B}_I\mathsf{B}_I^\top)$	0.51	0.52	0.57	0.75	0.89	0.82	0.87
$\lambda_{\max}(\mathsf{B}_I\mathsf{B}_I^\top)$	15.5	5.72	2.01	1.32	1.14	1.30	1.20
$\kappa(\mathsf{B}_I\mathsf{B}_I^\top)$	30.4	11.4	3.55	1.77	1.28	1.58	1.38

15.4 Combining Mortar Non-penetration with FETI Interconnecting

To plug our results into TFETI or TBETI, let us decompose each body Ω^1, Ω^2 into the subdomains Ω_i^1, Ω_j^2 of the diameter less or equal to H in a way which complies with the triangulations $\mathscr{T}^k$, $k \in \{1, 2\}$, and use the triangulation to discretize separately the subdomains. The procedure introduces several local degrees of freedom for the displacements associated with the nodes on the interfaces $\overline{\Omega}_i^k \cap \overline{\Omega}_j^k$. To simplify the reference to the above discussions, let us first consider only the variables associated with the nodes on Γ_C and denote by $\mathbf{x}_C$ and $\widetilde{\mathbf{x}}_C$ the variables associated with the

problem without the decomposition considered above and the decomposed problem, respectively. Using a suitable numbering of both sets of variables, we can define the matrix $\mathsf{L}_C \in \mathbb{R}^{\widetilde{n}_C \times n_C}$ which assigns to each vector $\mathbf{x}_C \in \mathbb{R}^{n_C}$ of global variables the vector $\widetilde{\mathbf{x}}_C \in \mathbb{R}^{\widetilde{n}_C}$ of the variables associated with the decomposed problem so that

$$\widetilde{\mathbf{x}}_C = \mathsf{L}_C \mathbf{x}_C, \quad \mathsf{L}_C = \operatorname{diag}(\mathbf{e}_1, \cdots, \mathbf{e}_{n_C}), \quad \mathbf{e}_i = [1, \ldots, 1]^T \in \mathbb{R}^{m_i}.$$

Thus each component x_i associated with a node which belongs to m_i faces of the subdomains of Ω^k is assigned the vector $\widetilde{\mathbf{x}}_i = x_i \mathbf{e}_i \in \mathbb{R}^{m_i}$. To test the non-penetration of the decomposed problem, let us define the matrices

$$\widetilde{\mathsf{L}}_C = \operatorname{diag}(\|\mathbf{e}_1\|_1^{-1} \mathbf{e}_1, \cdots, \|\mathbf{e}_{n_C}\|_1^{-1} \mathbf{e}_{n_C}), \quad \widetilde{\mathsf{B}}_N = \mathsf{B}_N \widetilde{\mathsf{L}}_C^T, \tag{15.19}$$

so that

$$\widetilde{\mathsf{B}}_N \widetilde{\mathbf{x}}_C = \mathsf{B}_N \widetilde{\mathsf{L}}_C^T \widetilde{\mathbf{x}}_C = \mathsf{B}_N \widetilde{\mathsf{L}}_C^T \mathsf{L}_C \mathbf{x}_C = \mathsf{B}_N \mathbf{x}_C.$$

To enforce the gluing of the subdomains as in Chap. 11, notice that the displacement $\widetilde{\mathbf{y}} \in \operatorname{Im}\mathsf{L}_C$ if and only if $\mathbf{y}_i = y_i \mathbf{e}_i$, where $\mathbf{y}_i$ is a block of $\mathbf{y}$ induced by the block structure of $\operatorname{Im}\mathsf{L}_C$. It follows that $\operatorname{Im}\mathsf{L}_C$ comprises the displacements which correspond to the displacements of the glued subdomains of Ω^k, $k \in \{2, 1\}$, so that the part $\widetilde{\mathsf{B}}_G$ of the "gluing" matrix that acts on the variables associated with Γ_C satisfies

$$\operatorname{Im}\widetilde{\mathsf{L}}_C = \operatorname{Im}\mathsf{L}_C = \operatorname{Ker}\widetilde{\mathsf{B}}_G, \quad \widetilde{\mathsf{B}}_G = \begin{bmatrix} \widetilde{\mathsf{B}}_G^1 & \mathsf{O} \\ \mathsf{O} & \widetilde{\mathsf{B}}_G^2 \end{bmatrix}.$$

The blocks $\widetilde{\mathsf{B}}_G^i$ can be reordered to be block diagonal, with the nonzero blocks comprising the row vectors of an orthonormal basis of $\operatorname{Ker} \mathbf{e}_i^T \subset \mathbb{R}^{m_i}$ so that

$$\widetilde{\mathsf{L}}_C^T \widetilde{\mathsf{B}}_G^T = \mathsf{O}$$

and

$$\widetilde{\mathsf{B}}_N \widetilde{\mathsf{B}}_G^T = \mathsf{B}_N \widetilde{\mathsf{L}}_C^T \widetilde{\mathsf{B}}_G^T = \mathsf{O}. \tag{15.20}$$

Let us now return to the discretization of our two body contact problems, so that $\mathbf{x} \in \mathbb{R}^n$ denotes the vector of coordinates of $\mathbf{v}_h \in V_h$. The matrices describing the non-penetration and gluing of variables that correspond to the nodes on Γ_C can be obtained by padding $\widetilde{\mathsf{B}}_N$ and $\widetilde{\mathsf{B}}_G$ with zeros. If we add the matrices $\widetilde{\mathsf{B}}_R$ with orthonormal rows which enforce the gluing of the remaining variables and enforcing the Dirichlet conditions, we get the constraint matrix

$$\mathsf{B}_I = \begin{bmatrix} \widetilde{\mathsf{B}}_N & \mathsf{O} \end{bmatrix}, \quad \mathsf{B}_E = \begin{bmatrix} \widetilde{\mathsf{B}}_G & \mathsf{O} \\ \mathsf{O} & \widetilde{\mathsf{B}}_R \end{bmatrix}, \quad \mathsf{B} = \begin{bmatrix} \widetilde{\mathsf{B}}_N & \mathsf{O} \\ \widetilde{\mathsf{B}}_G & \mathsf{O} \\ \mathsf{O} & \widetilde{\mathsf{B}}_R \end{bmatrix} = \begin{bmatrix} \mathsf{B}_I \\ \mathsf{B}_E \end{bmatrix},$$

and

$$\mathsf{B}\mathsf{B}^T = \operatorname{diag}(\mathsf{B}_I\mathsf{B}_I^T, \mathsf{B}_E\mathsf{B}_E^T) = \operatorname{diag}(\widetilde{\mathsf{B}}_N\widetilde{\mathsf{B}}_N^T, \mathsf{I}). \tag{15.21}$$

Using (15.19) and assuming that B_N has m_C rows, we get for any $\boldsymbol{\lambda} \in \mathbb{R}^{m_C}$

$$\|\widetilde{\mathsf{B}}_N^T\boldsymbol{\lambda}\|^2 = \boldsymbol{\lambda}^T\mathsf{B}_N\widetilde{\mathsf{L}}_C^T\widetilde{\mathsf{L}}_C\mathsf{B}_N^T\boldsymbol{\lambda} = \boldsymbol{\lambda}^T\mathsf{B}_N\Sigma\mathsf{B}_N^T\boldsymbol{\lambda} = \|\Sigma^{1/2}\mathsf{B}_N^T\boldsymbol{\lambda}\|^2,$$

where

$$\Sigma = \operatorname{diag}(\|\mathbf{e}_1\|_1^{-1}, \ldots, \|\mathbf{e}_{n_C}\|_1^{-1}).$$

To give bounds on the singular values of $\widetilde{B}$, recall that $\|\mathbf{e}_i\|_1$ denotes the number of subdomains the boundaries of which contain the node associated with the variable x_i and denote

$$s_{\max} = \max_{i=1,\ldots,n_C} \|\mathbf{e}_i\|_1 \geq 1.$$

Then

$$s_{\max}^{-1/2}\|\mathsf{B}_N^T\boldsymbol{\lambda}\| \leq \|\widetilde{\mathsf{B}}_N^T\boldsymbol{\lambda}\| \leq \|\mathsf{B}_N^T\boldsymbol{\lambda}\|. \tag{15.22}$$

Since the rows of B_N and B_{E*} are normalized and orthonormalized, respectively, we can combine (15.21) and (15.22) to get that for any $\boldsymbol{\lambda}$

$$s_{\max}^{-1}\|\mathsf{B}_N^T\boldsymbol{\lambda}_I\|^2 + \|\boldsymbol{\lambda}_E\|^2 \leq \|\mathsf{B}^T\boldsymbol{\lambda}\|^2 \leq \|\mathsf{B}_N^T\boldsymbol{\lambda}_I\|^2 + \|\boldsymbol{\lambda}_E\|^2.$$

Under the assumptions of Corollary 15.1, we get

$$\frac{1}{7\overline{s}_{\max}}\|\boldsymbol{\lambda}\|^2 \leq \|\mathsf{B}^T\boldsymbol{\lambda}\|^2 \leq (2+\overline{k})\|\boldsymbol{\lambda}\|^2. \tag{15.23}$$

Now we are ready to formulate the main result on the conditioning of the constraints arising in the solution of frictionless contact problems by the TFETI method with the non-penetration enforced by biorthogonal mortars.

Proposition 15.1 *Let the multibody contact problem introduced in Sect. 11.4 be discretized by linear finite elements reproducing constants with the consistent mortar discretization of the conditions of non-penetration. Let* $\overline{\Gamma}_C^{ij} \cap \overline{\Gamma}_C^{kl} = \emptyset$ *and* $\chi(\overline{\Gamma}_C^{ij}) \cap \chi(\overline{\Gamma}_C^{kl}) = \emptyset$ *for any* (i, j), (k, l) *that belong to the coupling set* $\mathscr{S}$. *Then the constraint matrix* B *can be formed in such a way that*

$$\frac{1}{7s_{\max}} < \lambda_{\min}\left(\mathsf{B}\mathsf{B}^T\right) \leq \|\mathsf{B}\|^2 \leq 2 + \overline{k}, \tag{15.24}$$

where $\overline{k}$ *denotes the cover number of the mortar discretization and* $s_{\max}$ *denotes the maximum number of the subdomains of one body that share a point on* Γ_C.

Proof First notice that inequalities (15.23) are valid for any couple of bodies that can come into contact. It follows that if we denote by $\mathsf{B}_{\mathscr{I}_{ij}*}$ the block of rows of B that enforce the gluing or non-penetration of bodies (or domains) on the interface Γ_C^{ij}, $(i, j) \in \mathscr{S}$, then for each (i, j), $(j, \ell) \in \mathscr{S}$

$$\frac{1}{7s_{\max}} < \lambda_{\min}\left(\mathsf{B}_{\mathscr{I}_{ij}*}\mathsf{B}^T_{\mathscr{I}_{ij}*}\right) \le \|\mathsf{B}_{\mathscr{I}_{ij}*}\|^2 \le 2 + \bar{k}.$$

Moreover, the assumption guarantees that for each $(i, j), (k, \ell) \in \mathscr{S}$, $(i, j) \neq (k, l)$, we have

$$\mathsf{B}_{\mathscr{I}_{ij}*}\mathsf{B}^T_{\mathscr{I}_{kl}*} = \mathsf{O}.$$

Since the blocks $\mathsf{B}_{\mathscr{I}_{ij}*}$, $(i, j) \in \mathscr{S}$ are orthogonal to the block $\widetilde{\mathsf{B}}_R$ of the orthonormal rows of B that define the equality constraints which do not interfere with the contact interface, it follows that the estimates (15.23) are valid also for the multibody contact problems that satisfy the assumptions of the proposition. □

Proposition 15.1 shows that the matrix F arising from the mortar discretization of the non-penetration conditions can be assembled so that it satisfies the assumptions of Theorem 11.1 which established the H/h bound on the regular condition number $\overline{\kappa}(\mathsf{PFP})$ with B generated by means of node-to-node constraints. It follows that Theorem 11.2 on optimality of TFETI for the solution of frictionless problems is valid also for the non-penetration enforced by biorthogonal mortars.

15.5 Numerical Experiments

We have implemented the mortar approximation of the non-penetration conditions in MATLAB using our MatSol numerical library [5] and examined the conditioning of both raint and Schur complement matrices on three benchmarks.

15.5.1 3D Hertz Problem with Decomposition

We shall use the benchmark depicted in Fig. 15.2 (left) that was used in Vlach et al. [4]. The upper body Ω_2 is pressed down on the upper face $x_3 = 10$ by the boundary traction $(0, 0, -2e3)$ [MPa]. The material coefficients were the same for both bodies, $\nu = 0.3$ and $E = 210$ [GPa]. The symmetry conditions are imposed on the boundaries $x_1 = 0$ and $x_2 = 0$; the lower body is fixed vertically along the bottom face $x_3 = -10$.

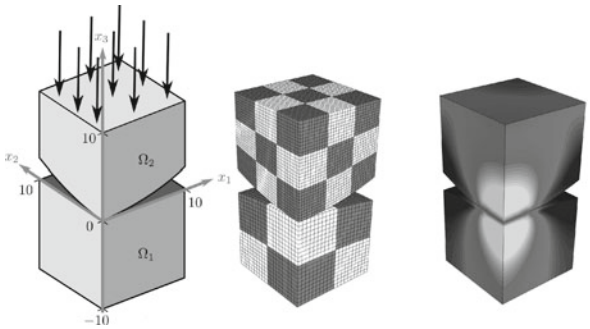

Fig. 15.2 3d Herz example – setting (*left*); decomposition case (*middle*); von Mises stress (*right*)

To illustrate our comments on the application of the theory in the analysis of the domain decomposition methods, we carried out the computations with each body either undecomposed or decomposed into $3 \times 3 \times 3$ or $2 \times 2 \times 2$ subdomains, depending on $h_{\text{slave}}/h_{\text{master}}$. The decomposition for $h_{\text{slave}}/h_{\text{master}} = 2/3$ is in Fig. 15.2 (middle).

Table 15.2 3D Hertz problem with domain decomposition

$h_{\text{slave}}/h_{\text{master}}$	1/6	1/3	2/3	1	3/2	3	6
$\lambda_{\min}(\mathsf{B}_I\mathsf{B}_I^\top)$	0.49	0.40	0.43	0.73	0.86	0.83	0.88
$\lambda_{\max}(\mathsf{B}_I\mathsf{B}_I^\top)$	19.2	7.00	2.30	1.32	1.16	1.31	1.20
$\kappa(\mathsf{B}_I\mathsf{B}_I^\top)$	42.8	17.5	5.36	1.82	1.35	1.57	1.37
$\lambda_{\min}(\mathsf{F}\|\mathsf{Ker}\mathsf{G})$	1.6e-6	2.4e-6	2.8e-6	3.8e-6	2.7e-6	2.2e-6	1.4e-6
$\lambda_{\max}(\mathsf{F}\|\mathsf{Ker}\mathsf{G})$	1.1e-4	1.1e-4	1.1e-4	1.1e-4	9.9e-5	9.9e-5	9.9e-5
$\kappa(\mathsf{F}\|\mathsf{Ker}\mathsf{G})$	72.8	46.7	40.7	29.9	37.2	44.0	72.7

The results are in Table 15.2. We can see that the decomposition hurts the conditioning of the dual Schur complement F rather mildly, being obviously affected more by the irregular decomposition which is far from the one required by the theory of domain decomposition reported in Chap. 11. In the right part of Fig. 15.2, the distribution of von Mises stress in the deformed state is depicted. The number of matrix–vector multiplications that were used to get the solution ranged from 72 to 240.

15.6 Comments and References

Mortars were introduced into domain decomposition methods by Maday, Mavripilis, and Patera [6] with the aim to admit non-matching discretization of subdomains. The mortar approximation of contact conditions was used by many researchers including Puso [7], Puso and Laursen [8], Wriggers [9], Dickopf and Krause [10], and Chernov et al. [11], and was enhanced into the efficient algorithms for contact problems as in Wohlmuth and Krause [12].

The biorthogonal mortars, which were introduced by Wohlmuth [1], turned out to be the effective tool for solving problems by means of domain decomposition. Our exposition is based on the variationally consistent approximation of contact conditions by biorthogonal mortars that can be found in the seminal paper by Wohlmuth [2]. See also Popp et al. [3]. Her presentation includes friction and the discretized problems are solved by means of multigrid in the framework of the nonsmooth Newton methods.

The estimates for two bodies without the decomposition are due to Vlach (see Vlach et al. [4]).

References

1. Wohlmuth, B.I.: Discretization Methods and Iterative Solvers Based on Domain Decomposition. Springer, Berlin (2001)
2. Wohlmuth, B.I.: Variationally consistent discretization scheme and numerical algorithms for contact problems. Acta Numerica **20**, 569–734 (2011)
3. Popp, A., Seitza, A., Geeb, M.W., Wall, W.A.: Improved robustness and consistency of 3D contact algorithms based on a dual mortar approach. Comput. Methods Appl. Mech. Eng. **264**, 67–80 (2013)
4. Vlach, O., Dostál, Z., Kozubek, T.: On conditioning the constraints arising from variationally consistent discretization of contact problems and duality based solvers. Comput. Methods Appl. Math. **15**(2), 221–231 (2015)
5. Kozubek, T., Markopoulos, A., Brzobohatý, T., Kučera, R., Vondrák, V., Dostál, Z.: MatSol–MATLAB efficient solvers for problems in engineering. http://industry.it4i.cz/en/products/matsol/ (2015)
6. Maday, Y., Mavriplis, C., Patera, A.T.: Nonconforming mortar element methods: application to spectral discretizations. In: Chan, T. (ed.) Domain Decomposition Methods, pp. 392–418. SIAM, Philadelphia (1989)
7. Puso, M.: A 3D mortar method for solid mechanics. Int. J. Numer. Methods Eng. **59**, 315–336 (2004)
8. Puso, M., Laursen, T.: A mortar segment-to-segment contact method for large deformation solid mechanics. Comput. Methods Appl. Mech. Eng. **193**, 601–629 (2004)
9. Wriggers, P.: Contact Mechanics. Springer, Berlin (2005)
10. Dickopf, T., Krause, R.: Efficient simulation of multi-body contact problems on contact geometries: a flexible decomposition approach using constrained optimization. Int. J. Numer. Methods Eng. **77**(13), 1834–1862 (2009)
11. Chernov, A., Maischak, M., Stephan, E.P.: hp-mortar boundary element method for two-body contact problems with friction. Math. Methods Appl. Sci. **31**, 2029–2054 (2008)
12. Wohlmuth, B.I., Krause, R.: Monotone methods on nonmatching grids for nonlinear contact problems. SIAM J. Sci. Comput. **25**, 324–347 (2003)

Chapter 16
Preconditioning and Scaling

So far, the scalability results assumed that the bodies and subdomains involved in a problem are made of the same material and that the stiffness matrices of the subdomains are reasonable conditioned. If this is not the case, we can try to improve the rate of convergence of the solvers by preconditioning.

Here we present some results that can be used to improve the convergence of the algorithms for the solution of QP or QCQP problems arising from the application of TFETI to the solution of contact problems. We are interested especially in the methods which not only improve the condition number of the Hessian of the cost function but also preserve the structure of the inequality constraints, so that they affect both the linear and nonlinear steps.

We first consider a preconditioning which satisfies the above requirements and can reduce or eliminate the effect of varying coefficients. Our development uses the structure of the matrices F, B that were introduced in Chap. 11 and results concerning the reduction of upper bounds of the condition number by a diagonal scaling. We show that the proposed preconditioning guarantees the bounds on the regular condition number of preconditioned systems that are independent of the coefficients provided the subdomains are homogeneous and each node is involved in at most one inequality constraint. The preconditioning can be implemented in such a way that it preserves the bound constraints and affects both linear and nonlinear steps. If some node is involved in more than one inequality then it can be used for the preconditioning in face (see Sect. 8.6). A simplified diagonal stiffness scaling, which preserves the bound constraints, is shown to reduce the ill-conditioning caused by varying coefficients in more general case.

Then we discuss the adaptation of standard Dirichlet and lumped preconditioners to solving contact problems and their implementation into the preconditioning in face of the MPRGP and MPGP algorithms.

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3_16

16.1 Reorthogonalization-Based Preconditioning

We shall start with the preconditioning of the matrix

$$\mathsf{F} = \mathsf{B}^T \mathsf{K}^+ \mathsf{B}, \qquad \mathsf{K} = \mathrm{diag}(\mathsf{K}_1, \cdots, \mathsf{K}^s),$$

arising from the application of FETI domain decomposition methods to the solution of frictionless contact problems with matching grids and homogeneous subdomains made of isotropic material, so that their stiffness matrices depend linearly on their modulus of elasticity. We assume that the matrices B and K were obtained by the procedure described in Chap. 11. Notice that the constraint matrix B can be reordered so that the reordered $\widetilde{\mathsf{B}}$ is block diagonal with small blocks.

The structure of F, B and Lemma 10.2 suggests that we can try to improve the conditioning of F by first improving the regular condition number of K by the diagonal scaling to get

$$\widetilde{\mathsf{K}} = \Delta^{-1/2} \mathsf{K} \Delta^{-1/2}, \quad \Delta = \mathrm{diag}(k_{11}, \cdots, k_{nn}),$$

and

$$\mathsf{F} = \mathsf{B}\mathsf{K}^+\mathsf{B}^T = \mathsf{B}\Delta^{-1/2}\widetilde{\mathsf{K}}^+\Delta^{-1/2}\mathsf{B}^T,$$

so that $\widetilde{\mathsf{K}}$ does not depend on the Young coefficients E_i, $i = 1, \ldots, s$, and then determine T by attempting to make

$$\widetilde{\mathsf{B}} = \mathsf{T}\mathsf{B}\Delta^{-1/2}$$

as close to a matrix with orthonormal rows as possible. If each block of $\widetilde{B}$ contains at most one inequality, we can use T arising from the properly ordered Gram–Schmidt orthogonalization of $\mathsf{B}\Delta^{-1/2}$ to get

$$\mathsf{T}\mathsf{B}\Delta^{-1/2}(\mathsf{T}\mathsf{B}\Delta^{-1/2})^\top = \mathsf{I}. \tag{16.1}$$

Notice that $\widetilde{\mathsf{B}}$ is a reordered block diagonal matrix, so it is not expensive to get $\mathsf{B}\Delta^{-1/2}(\mathsf{B}\Delta^{-1/2})^T$, its Cholesky factorization $\mathsf{L}\mathsf{L}^T = \mathsf{B}\Delta^{-1}\mathsf{B}^T$, and the action of $\mathsf{T} = \mathsf{L}^{-1}$. The effect of the reortogonalization-based preconditioning is formulated in the following proposition.

Proposition 16.1 *Let the domain of a contact problem be decomposed into homogeneous subdomains of the diameter H and discretized by the shape regular elements with the diameter h. Let the discretization be quasi-uniform with shape regular elements and let the subdomains be homogeneous and isotropic.*

Then the regular condition number κ satisfies

$$\overline{\kappa}(\mathsf{TFT}^T) = \overline{\kappa}(\widetilde{\mathsf{B}}\widetilde{\mathsf{K}}^{\dagger}\widetilde{\mathsf{B}}^T) \le C\frac{H}{h}, \tag{16.2}$$

where C is independent of H, h, and Young's modulus.

Proof Neither $\widetilde{\mathsf{K}}$ nor the conditioning of $\widetilde{\mathsf{B}}$ depend on Young's modulus and

$$\mathsf{TFT}^T = \mathsf{TBK}^{\dagger}\mathsf{B}^T\mathsf{T}^T = \widetilde{\mathsf{B}}\widetilde{\mathsf{K}}^{\dagger}\widetilde{\mathsf{B}}^T.$$

The rest follows by Theorem 11.1. □

We can plug the reorthogonalization-based preconditioning into problem (11.40)

$$\min \overline{\theta}(\boldsymbol{\lambda}) \quad \text{s.t.} \quad \mathsf{G}\boldsymbol{\lambda} = \mathbf{o} \quad \text{and} \quad \boldsymbol{\lambda}_{\mathscr{I}} \ge \boldsymbol{\ell}_{\mathscr{I}}$$

by substituting $\boldsymbol{\lambda} = \mathsf{T}^T\boldsymbol{\mu}$ to get

$$\min \overline{\theta}(\mathsf{T}^T\boldsymbol{\mu}) \quad \text{s.t.} \quad \mathsf{GT}^T\boldsymbol{\mu} = \mathbf{o} \quad \text{and} \quad [\mathsf{T}^T\boldsymbol{\mu}]_{\mathscr{I}} \ge \boldsymbol{\ell}_{\mathscr{I}}. \tag{16.3}$$

Since the reorthogonalization-based preconditioning turns in general case the bound constraints into more general inequality constraints, it follows that we can use this preconditioning only as preconditioning in face. However, if each node is involved in at most one inequality constraint, as in the two-body problem with a smooth contact interface, i.e., if each block of $\widetilde{\mathsf{B}}$ has at most one row corresponding to the inequality constraint, then it is possible to preserve the bound inequality constraints. It is just enough to start the orthogonalization of each block of $\widetilde{\mathsf{B}}$ with the rows corresponding to the equality constraints to get

$$\widetilde{\mathsf{B}} = \begin{bmatrix} \mathsf{T}_E & \mathsf{O} \\ \mathsf{T}_I & \Sigma \end{bmatrix} \begin{bmatrix} \mathsf{B}_E \\ \mathsf{B}_I \end{bmatrix}, \tag{16.4}$$

so that

$$\min \overline{\theta}(\mathsf{T}^T\boldsymbol{\mu}) \quad \text{s.t.} \quad \mathsf{GT}^T\boldsymbol{\mu} = \mathbf{o}, \quad \boldsymbol{\mu}_{\mathscr{I}} \ge \Sigma^{-1}\boldsymbol{\ell}_{\mathscr{I}}, \quad \boldsymbol{\lambda} = \mathsf{T}^T\boldsymbol{\mu}, \tag{16.5}$$

where Σ is a diagonal matrix with positive entries. Since such transformation preserves the bound constraints, it follows that the preconditioning by reorthogonalization can be applied to discretized dual problem (16.3) and improve the bounds on the rate of convergence of the algorithms like MPRGP, MPGP, SMALBE, or SMALSE introduced in Part II.

16.2 Renormalization-Based Stiffness Scaling

Let us now consider a general bound and equality constrained problems (11.36) arising from the discretization of a contact problem with nonsmooth interface or which admits nodes on the contact interface which belong to more than two bodies, so that the reorthogonalization-based preconditioning changes the bound constraints into more general inequality constraints (16.3). To improve the conditioning of the non-linear steps, we can try to use Δ to find a diagonal matrix

$$\Sigma = \mathrm{diag}[\sigma_1, \ldots, \sigma_m]$$

with positive diagonal entries such that the regular condition number of the scaled matrix $\Sigma\mathsf{F}\Sigma$ is smaller than that of F. Such scaling can be considered as modification of the constraints to

$$\Sigma_{\mathscr{I}}\mathsf{B}_{\mathscr{I}}\mathbf{v} \le \Sigma_{\mathscr{I}}\mathbf{g}_{\mathscr{I}}, \quad \Sigma_{\mathscr{E}}\mathsf{B}_{\mathscr{E}}\mathbf{v} = \mathbf{o}, \quad \text{and} \quad \Sigma = \mathrm{diag}[\Sigma_{\mathscr{I}}, \Sigma_{\mathscr{E}}],$$

where $\Sigma_{\mathscr{I}}$ and $\Sigma_{\mathscr{E}}$ are the diagonal blocks of Σ that correspond to $\mathsf{B}_{\mathscr{I}}$ and $\mathsf{B}_{\mathscr{E}}$, respectively.

As above, we shall first improve the regular condition number of K by the diagonal scaling to get

$$\widetilde{\mathsf{K}} = \Delta^{-1/2}\mathsf{K}\Delta^{-1/2}$$

and

$$\mathsf{F} = \mathsf{B}\mathsf{K}^{\dagger}\mathsf{B}^T = \mathsf{B}\Delta^{-1/2}\widetilde{\mathsf{K}}^{\dagger}\Delta^{-1/2}\mathsf{B}^T,$$

and then determine Σ by attempting to make $\widetilde{\mathsf{B}} = \Sigma\mathsf{B}\Delta^{-1/2}$ as close to a matrix with orthonormal rows as possible, i.e.,

$$\Sigma\mathsf{B}\Delta^{-1/2}(\Sigma\mathsf{B}\Delta^{-1/2})^T \approx \mathsf{I}. \tag{16.6}$$

Probably the simplest idea is to require that (16.6) is satisfied by the diagonal entries. Denoting by $\mathbf{b}_i$ the i-th row of B, we get for the diagonal entries σ_i of Σ

$$\sigma_i\mathbf{b}_i\Delta^{-1/2}(\sigma_i\mathbf{b}_i\Delta^{-1/2})^T \quad \text{and} \quad \sigma_i = 1/\sqrt{\mathbf{b}_i\Delta^{-1}\mathbf{b}_i^T}. \tag{16.7}$$

Unfortunately, it turns out that the scaling is not sufficient to get an optimality result similar to Proposition 16.1, though numerical experiments indicate efficiency of the diagonal scaling. The problem is in the constraints which enforce the continuity of the solution in the corners of the subdomains with different coefficients. The following proposition on scalar problem partly explains both observations.

Proposition 16.2 *Let a 2D Poisson problem on square be decomposed into homogeneous subdomains and discretized by finite elements using a regular matching*

grids. Let the continuity of subdomains be enforced across the faces by the equations $x_i - x_j = 0$ *and in the corners by the equations*

$$x_i + x_j - x_k - x_\ell = 0, \quad x_i - x_j = 0, \quad \text{and} \quad x_k - x_\ell = 0.$$

Let the constraint matrix $\widetilde{\mathsf{B}} = \Sigma \mathsf{B} \Delta^{-1/2}$ *and the vectors of Lagrange multipliers* $\boldsymbol{\lambda} \in \mathbb{R}^m$ *be decomposed into the blocks* $\widetilde{\mathsf{B}}_c$ *and* $\boldsymbol{\lambda}_c \in \mathbb{R}^{m_c}$ *associated with the corner constraints and* $\widetilde{\mathsf{B}}_e$ *and* $\boldsymbol{\lambda}_e \in \mathbb{R}^{m_e}$ *associated with the edge constraints,* $m = m_c + m_e$*, so that*

$$\widetilde{\mathsf{B}}^T \boldsymbol{\lambda} = \widetilde{\mathsf{B}}_c^T \boldsymbol{\lambda}_c + \widetilde{\mathsf{B}}_e^T \boldsymbol{\lambda}_e. \tag{16.8}$$

Then there is $0 < C_1 \le 1$ *such that*

$$C_1 \|\boldsymbol{\lambda}_c\|^2 \le \|\widetilde{\mathsf{B}}_c^T \boldsymbol{\lambda}_c\|^2 \le 2\|\boldsymbol{\lambda}_c\|^2 \quad \text{and} \quad \|\widetilde{\mathsf{B}}_e^T \boldsymbol{\lambda}_e\|^2 = \|\boldsymbol{\lambda}_e\|^2. \tag{16.9}$$

Moreover, there there are subspaces $V_c \subseteq \mathbb{R}^{m_c}$ *and* $W_c \subseteq \mathbb{R}^{m_c}$ *of the dimension* n_c *and* $m_c - n_c$*, where* n_c *is the number of corners, such that* $\mathbb{R}^{m_c} = V_c \oplus W_c$ *and for any* $\boldsymbol{\lambda}_c \in V_c$ *and* $\mu_c \in W_c$

$$\boldsymbol{\lambda}_c^T \mu_c = 0, \quad C_1 \|\boldsymbol{\lambda}_c\|^2 \le \|\widetilde{\mathsf{B}}_c^T \boldsymbol{\lambda}_c\|^2 \le \|\boldsymbol{\lambda}_c\|^2, \quad \text{and} \quad \|\mu_c\|^2 \le \|\widetilde{\mathsf{B}}_c^T \mu_c\|^2 \le 2\|\mu_c\|^2. \tag{16.10}$$

Proof Let us denote by $\widetilde{\mathsf{B}}_c^{ijk\ell}$ the 3×4 block of $\widetilde{\mathsf{B}}_c$ associated with variables v_i, v_j, v_k, v_ℓ which are glued by the corresponding three rows of $\widetilde{\mathsf{B}}_c$. Denoting the parts of these rows $\widetilde{\mathbf{b}}_1, \widetilde{\mathbf{b}}_2, \widetilde{\mathbf{b}}_3$, we get

$$\widetilde{\mathsf{B}}_c^{ijk\ell} = \begin{bmatrix} \widetilde{\mathbf{b}}_1 \\ \widetilde{\mathbf{b}}_2 \\ \widetilde{\mathbf{b}}_3 \end{bmatrix}, \quad \widetilde{\mathsf{B}}_c^{ijk\ell} (\widetilde{\mathsf{B}}_c^{ijk\ell})^T = \begin{bmatrix} 1 & \widetilde{\mathbf{b}}_1^T \widetilde{\mathbf{b}}_2 & \widetilde{\mathbf{b}}_1^T \widetilde{\mathbf{b}}_3 \\ \widetilde{\mathbf{b}}_1^T \widetilde{\mathbf{b}}_2 & 1 & 0 \\ \widetilde{\mathbf{b}}_1^T \widetilde{\mathbf{b}}_3 & 0 & 1 \end{bmatrix}.$$

The direct computation using the properties of the eigenvalues of SPS matrices shows that the eigenvalues of $\widetilde{\mathsf{B}}_c^{ijk\ell} (\widetilde{\mathsf{B}}_c^{ijk\ell})^T$ are given by

$$\mu_1 = 1 - \sqrt{(\widetilde{\mathbf{b}}_1^T \widetilde{\mathbf{b}}_2)^2 + (\widetilde{\mathbf{b}}_1^T \widetilde{\mathbf{b}}_3)^2} > 0, \ \mu_2 = 1, \ \mu_3 = 1 + \sqrt{(\widetilde{\mathbf{b}}_1^T \widetilde{\mathbf{b}}_2)^2 + (\widetilde{\mathbf{b}}_1^T \widetilde{\mathbf{b}}_3)^2} < 2.$$

It follows that we can take for C_1 the smallest eigenvalue of the corner blocks. The rest is easy as the rows of B_e are orthonormal and orthogonal to the rows of B_c. □

Remark 16.1 Proposition 16.2 indicates that the effect of varying coefficients can be reduced by the scaling to a relatively small subspace which does not contain dominating components of the solution, so that it can improve the convergence of the iterative contact problems. Notice that the scaling is sufficient to get rid of the effect of varying coefficients for 2D problems solved by FETI-DP. Nevetherless, the results indicate that the effect of scaling is limited, at least as far as the non-redundant coefficients are concerned.

16.3 Lumped and Dirichlet Preconditioners in Face

There are several well established preconditioners that can be used to improve the conditioning of the Hessian PFP of the dual function. We can try to adapt them to improve the rate of convergence for the solution of auxiliary linear problems by preconditioning in face (see Sect. 8.6), but we shall see that the effect of adapted preconditioners is supported rather by the intuition than by the theory.

The *lumped preconditioner* was introduced by Farhat and Roux in their seminal paper on FETI [1]. It is very cheap and often outperforms more sophisticated preconditioners. For the linear problems solved by TFETI, it is given by

$$\mathsf{M} = \mathsf{BKB}^T = \mathsf{B}_{*\mathscr{B}}\mathsf{K}_{\mathscr{B}\mathscr{B}}\mathsf{B}_{*\mathscr{B}}^T, \tag{16.11}$$

where $\mathscr{B}$ denotes the set of all indices of the nodal variables that are associated with the nodes on the boundary of the subdomains.

The *Dirichlet preconditioner* was introduced by Farhat, Mandel, and Roux in their another important paper on FETI [2]. For the linear problems solved by TFETI, it is given by

$$\mathsf{M} = \mathsf{B}_{*\mathscr{B}}\mathsf{SB}_{*\mathscr{B}}^T, \tag{16.12}$$

The adaptation for the solution of contact problems is not obvious. Taking into account that the Hessian of the dual cost function $\overline{\theta}_\varrho$ in Sect. 11.8 reads $\mathsf{PFP} + \varrho\mathsf{Q}$, we can consider the adapted lumped preconditioner in the form

$$\mathsf{M} = \mathsf{BKB}^T = \mathsf{B}_{\mathscr{F}\mathscr{B}}\mathsf{K}_{\mathscr{B}\mathscr{B}}\mathsf{B}_{\mathscr{F}\mathscr{B}}^T + \varrho^{-1}\mathsf{Q}_{\mathscr{F}\mathscr{F}}, \tag{16.13}$$

where $\mathscr{F}$ denotes the indices of the variables associated with the free variables in face and the adapted Dirichlet preconditioner in the form

$$\mathsf{M} = \mathsf{B}_{\mathscr{F}\mathscr{B}}\mathsf{SB}_{\mathscr{F}\mathscr{B}}^T + \varrho^{-1}\mathsf{Q}_{\mathscr{F}\mathscr{F}}. \tag{16.14}$$

Though the second term is not a projector any more, our experience showed that the lumped preconditioner can modestly reduce the cost of the solution [3], [4]. Let us mention that replacing $\mathsf{Q}_{\mathscr{F}\mathscr{F}}$ by the projector onto the kernel of $\mathsf{B}_{\mathscr{F}\mathscr{B}}\mathsf{K}_{\mathscr{B}\mathscr{B}}\mathsf{B}_{\mathscr{F}\mathscr{B}}^T$ seems possible only at a rather high cost and that the preconditioning in face does not improve the performance of the nonlinear steps.

16.4 Numerical Experiments

The variants of scaling were incorporated into MatSol [5] and tested on 2D and 3D benchmarks.

16.4.1 3D Heterogeneous Beam

The domain of our first benchmark was a 3D beam of size 2×1 [m] divided into $6 \times 3 \times 3$ subdomains made of two materials with Young's modulus $E_1 = 1e3$ and $E_2 = \alpha 1e3$, $\alpha \in \{1, 10.100, 1000\}$. The Dirichlet boundary conditions were prescribed on $\Gamma_D = \{(0, y, z) : y \in \langle 0, 3\rangle, z \in \langle 0, 3\rangle\}$. Three different distributions of materials labeled by a, b, c that were used in our experiments are depicted in Fig. 16.1. The grey regions are filled with the material with Young's modulus E_2. The vertical traction 2000 [N] was applied on the upper face of the beam. The discretization resulted in a linear problem with 32,152 primal and 1,654 dual variables.

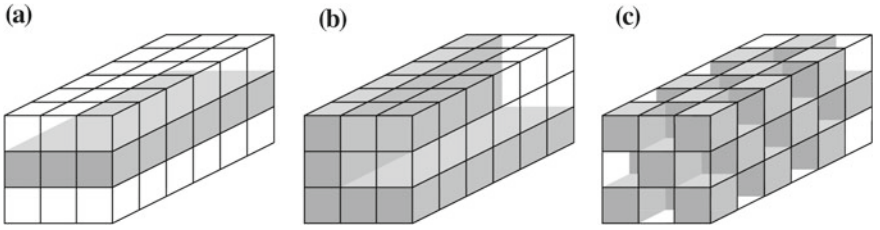

Fig. 16.1 3D heterogeneous beams – material distribution, E_2 grey

Table 16.1 3D beam—$\overline{\kappa}(PFP)$ for varying material distributions and preconditioning

E_2/E_1	I			$\mathsf{T}(\text{diag}\,\mathsf{K})$			$\Sigma(\text{diag}\,\mathsf{K})$		
	a	b	c	a	b	c	a	b	c
1	1.84e1	1.84e1	1.84e1	9.89e0	9.89e0	9.89e0	9.89e0	9.89e0	9.89e+00
10^{-1}	1.76e2	1.79e2	1.21e2	9.85e0	1.01e1	1.02e1	5.31e1	6.07e1	5.19e+01
10^{-2}	1.76e3	1.78e3	1.20e3	9.85e0	1.01e1	1.04e1	5.01e2	6.09e2	5.04e+02
10^{-3}	1.76e4	1.78e4	1.20e4	9.85e0	1.01e1	1.04e1	5.00e3	6.10e3	5.02e+03

The regular condition number of PFP for the varying combinations of material distributions a, b, c, material scales $E_2/E_1 \in \{1, 10, 100, 1000\}$, and preconditioning by I (no preconditioning), T or Σ are summarized in Table 16.1. We can see that the reorthogonalization-based preconditioning eliminates the effect of heterogeneous coefficients in agreement with the theory, while the renormalization-based scaling reduces the effect of jumps in the coefficients, but is far from eliminating it.

16.4.2 Contact Problem with Coulomb Friction

The effect of reorthogonalization-based preconditioning on a problem with at most one inequality associated with one node is documented by the analysis of the matrix with circular insets. The friction on the contact interfaces is described by Coulomb's law with the coefficient of friction $\Phi = 0.3$. The discretized problem was divided into subdomains in two ways using regular decomposition and Metis, see Fig. 16.2. The discretization and decomposition resulted in the problem with 9960 primal variables, 12(28) subdomains, 756(1308) dual variables, 36(84) equality constraints, and 384 box constraints.

The problem was resolved with varying E_2/E_1 and the reorthogonalization-based preconditioning T associated with the diagonal entries of K and S denoted respectively by $\mathsf{T}(\mathrm{diag}\mathsf{K})$ and $\mathsf{T}(\mathrm{diag}\mathsf{S})$. The results obtained by a regular and METIS discretization are reported separately. Young's modulus E_2 of the insets was parameterized as α times Young's modulus E_1 of the matrix. The resulting von Misses stress and the deformation are depicted in Fig. 16.3.

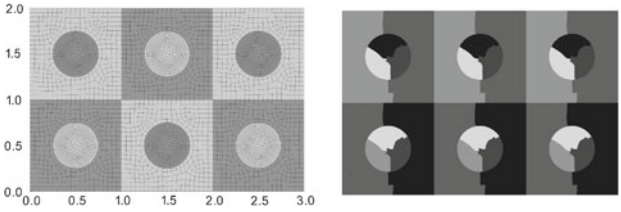

Fig. 16.2 Material with circular insets – regular (left) and Metis (right) decomposition

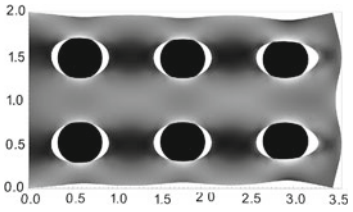

Fig. 16.3 Material with circular insets – deformation and the von Misses stress for $E_2/E_1 = 10000$

The effective condition number and the number of Hessian multiplications needed to achieve the relative error 1e-4 are presented in Table 16.2 (regular grid) and Table 16.3 (METIS grid). The regular condition number has been improved in all cases. Moreover, the number of iterations is nearly constant for regular decomposition in agreement with the theory as there are no corners on the material interface.

The computations with the Metis decomposition resulted in an increased number of Hessian multiplications. An obvious reason is the increased condition number due to the ill-conditioning of the constraint matrices.

Table 16.2 Material with circular insets – $\overline{\kappa}(\mathsf{PFP})$ / iterations (regular grid)

E_2/E_1	I	T(diag K)	T(diag S)
1	9.31e1 / 112	9.29e1 / 139	7.21e1 / 128
1e-1	1.26e2 / 129	6.20e1 / 145	5.27e1 / 144
1e-2	6.40e2 / 230	5.56e1 / 188	5.02e1 / 169
1e-3	6.26e3 / 371	5.50e1 / 148	4.99e1 / 128
1e-4	6.25e4 / 588	5.48e1 / 139	4.99e1 / 132

Table 16.3 Material with circular insets – $\overline{\kappa}(\mathsf{PFP})$ / iterations (Metis discretization)

E_2/E_1	I	T(diag K)	T(diag S)
1	3.85e2 / 203	3.40e2 / 287	3.22e2 / 269
1e-1	6.27e2 / 230	3.03e2 / 216	2.95e2 / 242
1e-2	6.23e3 / 345	2.95e2 / 249	2.89e2 / 207
1e-3	6.23e4 / 712	2.95e2 / 192	2.88e2 / 195
1e-4	6.23e5 / 972	2.94e2 / 251	2.88e2 / 200

16.5 Comments and References

The reorthogonalization based preconditioning and renormalization based scaling presented here appeared in Dostál et al. [6]. In spite of its simplicity, the renormalization based scaling is just one of the very few methods that are known to precondition the nonlinear steps that identify the contact interface (see also [7]). If there are no corners on the contact interface as in the problem solved in Sect. 16.3, we can enhance this preconditioning to get the rate of convergence independent of material coefficients. Otherwise we can use it as preconditioning in face (see Sect. 8.6).

The reorthogonalization-based scaling has a similar effect as the superlumped preconditioning developed by Rixen and Farhat (see [8] and [9]). See also Bhardway [10]. The results can be extended to the algorithms based on BETI (see Chap. 14).

The scaling is efficient mainly for problems with homogeneous domains of varying stiffness, but can be useful, at least in some special cases, also for heterogenous domains [11]. If all the subdomains have the same material coefficient, the renormalization based scaling reduces to the *multiplicity scaling*. The limits of the preconditioning effect that can be achieved by the diagonal scaling are indicated by the results of Greenbaum [12], Ainsworth, McLean, Tran [13], Forsythe and Strauss [14], van der Sluis [15]), or Bank and Scott [16].

References

1. Farhat, C., Roux, F.-X.: A method of finite element tearing and interconnecting and its parallel solution algorithm. Int. J. Numer. Methods Eng. **32**, 1205–1227 (1991)
2. Farhat, C., Mandel, J., Roux, F.-X.: Optimal convergence properties of the FETI domain decomposition method. Comput. Methods Appl. Mech. Eng. **115**, 365–385 (1994)
3. Dostál, Z., Gomes, F.A.M., Santos, S.A.: Solution of contact problems by FETI domain decomposition with natural coarse space projection. Comput. Methods Appl. Mech. Eng. **190**(13–14), 1611–1627 (2000)
4. Dostál, Z., Gomes, F.A.M., Santos, S.A.: Duality based domain decomposition with natural coarse space for variational inequalities. J. Comput. Appl. Math. **126**(1–2), 397–415 (2000)
5. Kozubek, T., Markopoulos, A., Brzobohatý, T., Kučera, R., Vondrák, V., Dostál, Z.: MatSol–MATLAB efficient solvers for problems in engineering. http://industry.it4i.cz/en/products/matsol/ (2015)
6. Dostál, Z., Kozubek, T., Vlach, O., Brzobohatý, T.: Reorthogonalization based stiffness preconditioning in FETI algorithms with applications to variational inequalities. Numer. Linear Algebr. Appl. **22**(6), 987–998 (2015)
7. Domorádová, M., Dostál, Z.: Projector preconditioning for partially bound constrained quadratic optimization. Numer. Linear Algebr. Appl. **14**(10), 791–806 (2007)
8. Rixen, D., Farhat, C.: Preconditioning the FETI method problems with intra- and inter-subdomain coefficient jumps. In Proceedings of the 9th International Conference on Domain Decomposition Methods, eds. P. Bjorstad, M. Espedal, and D. Keyes, Bergen, 472–479 (1997)
9. Rixen, D., Farhat, C.: A simple and efficient extension of a class of substructure based preconditioners to heterogeneous structural structural mechanics problems. Int. J. Numer. Methods Eng. **44**, 472–479 (1999)
10. Bhardwaj, M., Day, D., Farhat, C., Lesoinne, M., Pierson, K., Rixen, D.: Application of the FETI method to ASCI problems: scalability results on 1000 processors and discussion of highly heterogeneous problems. Int. J. Numer. Methods Eng. **47**, 513–535 (2000)
11. Kraus, J.K., Margenov, S.: Robust Algebraic Multilevel Methods and Algorithms. Radon Series on Computational and Applied Mathematics. De Gruiter, Berlin (2009)
12. Greenbaum, A.: Diagonal scaling of the Laplacian as preconditioners for other elliptic differential operators. SIAM J. Matrix Anal. Appl. **13**(3), 826–846 (1992)
13. Ainsworth, M., McLean, B., Tran, T.: Diagonal scaling of stiffness matrices in the Galerkin boundary element method. ANZIAM J. **42**(1), 141–150 (2000)
14. Forsythe, G.E., Strauss, E.G.: On best conditioned matrices. Proc. AMS - Am. Math. Soc. **6**, 340–345 (1955)
15. van der Sluis, A.: Condition numbers and equilibration of matrices. Numerische Mathematik **14**(1), 14–23 (1969)
16. Bank, R., Scott, R.: On the conditioning of finite element equations with highly refined meshes. SIAM J. Numer. Anal. **26**(6), 1383–1394 (1988)

Part IV
Other Applications and Parallel Implementation

Chapter 17
Contact with Plasticity

In the previous part of the book, we have presented the scalable algorithms for the solution of some contact problem which comprised elastic bodies, however, these algorithms are also useful for the solution of more general problems. Here, we shall indicate how to use the TFETI domain decomposition method for the solution of frictionless contact problems with elastic bodies that was introduced in Chap. 11 to the solution of contact problems with elasto-plastic bodies. Let us recall that plasticity is a time-dependent model of constitutive relations which takes into account the history of loading.

For simplicity, we shall restrict our attention to the two-body frictionless contact problem. We shall assume that the bodies occupy in a reference configuration the domains Ω^1, Ω^2 with the boundaries Γ^1, Γ^2 and their parts Γ_U^k, Γ_F^k, and Γ_C^k, $k = 1, 2$. The bodies are assumed to be fixed on Γ_U^k with a nonzero measure and can come in contact on Γ_C^k. The load is represented by the surface traction on Γ_F^k and by the volume forces defined on Ω^k, $k = 1, 2$. We shall assume that the constitutive relations are defined by associated elastoplasticity with the von Mises plastic criterion and the linear isotropic hardening law (see, e.g., Han and Reddy [1] or de Souza, Perić, and Owen [2]). More detailed description of the elasto-plastic initial-value constitutive model can be found, for example, in [3]. The weak formulation of the corresponding elasto-plastic problem can be found in [1].

Here, we start directly with the discretized elasto-plastic problem. For the sake of simplicity, we confine ourselves to the one-step problem formulated in displacements. Its solution is approximated by the iterates with the minimization of convex and smooth functional on a convex set in the inner loop. The discretized inner problem is a QP optimization problem with simple equality and inequality constraints which can be solved at the cost nearly proportional to the number of variables.

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3_17

17.1 Algebraic Formulation of Contact Problem for Elasto-Plastic Bodies

Let us start with the discretized formulation of the problem that arises from the application of the TFETI domain decomposition. We assume that each domain Ω^k is decomposed into the subdomains

$$\Omega_1^k, \Omega_2^k, \ldots, \Omega_{s_k}^k, \quad k = 1, 2,$$

and that each subdomain is assigned one index, so that the displacements can be described by the vector $\mathbf{v} \in \mathbb{R}^n$

$$\mathbf{v} = [\mathbf{v}_1^T, \ldots, \mathbf{v}_s^T]^T, \quad s = s_1 + s_2.$$

We define the subspace

$$V = \left\{ \mathbf{v} \in \mathbb{R}^n \ : \ \mathsf{B}_E \mathbf{v} = \mathbf{o} \right\}, \tag{17.1}$$

and the set of admissible displacement

$$\mathscr{K} = \left\{ \mathbf{v} \in \mathbb{R}^n \ : \ \mathsf{B}_E \mathbf{v} = \mathbf{o}, \ \mathsf{B}_I \mathbf{v} \le \mathbf{c}_I \right\}. \tag{17.2}$$

Here the matrix $\mathsf{B}_E \in \mathbb{R}^{m_E \times n}$ enforces the gluing of the subdomains and their fixing along the appropriate part of Γ^k, $k = 1, 2$. The inequality constraint matrix $\mathsf{B}_I \in \mathbb{R}^{m_I \times n}$ enforces the non-penetration condition on the contact interface. Notice that $\mathscr{K}$ is convex and closed.

The algebraic formulation of the contact elasto-plastic problem can be written as the following optimization problem [4]:

$$\text{Find } \mathbf{u} \in \mathscr{K} \ : \ J(\mathbf{u}) \le J(\mathbf{v}) \quad \forall \mathbf{v} \in \mathscr{K}, \tag{17.3}$$

where

$$J(\mathbf{v}) := \Psi(\mathbf{v}) - \mathbf{f}^T \mathbf{v}, \quad \mathbf{v} \in \mathbb{R}^n. \tag{17.4}$$

Here, the vector $\mathbf{f} \in \mathbb{R}^n$ represents the load consisting of the volume and surface forces and the initial stress state. The functional $\Psi : \mathbb{R}^n \to \mathbb{R}$ represents the inner energy, it is a potential to the nonlinear elasto-plastic operator $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$, i.e., $D\Psi(\mathbf{v}) = \mathbf{F}(\mathbf{v})$, $\mathbf{v} \in \mathbb{R}^n$. The function $\mathbf{F}$ is generally nonsmooth but Lipschitz continuous. It enables us to define a generalized derivative $\mathsf{K} : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ of $\mathbf{F}$ in the sense of Clark, i.e., $\mathsf{K}(\mathbf{v}) \in \partial \mathbf{F}(\mathbf{v})$, $\mathbf{v} \in \mathbb{R}^n$. Notice that $\mathsf{K}(\mathbf{v})$ is symmetric, block diagonal, and sparse matrix.

Problem (17.4) has a unique solution and can be equivalently written as the following variational inequality:

$$\text{Find } \mathbf{u} \in \mathscr{K} \ : \ F(\mathbf{u})^T(\mathbf{v}-\mathbf{u}) \geq \mathbf{f}^T(\mathbf{v}-\mathbf{u}) \quad \forall \mathbf{v} \in \mathscr{K}. \tag{17.5}$$

The properties of $\mathbf{F}$ and K can be found in Čermák and Sysala [3].

17.2 Semismooth Newton Method for Optimization Problem

Problem (17.3) contains two nonlinearities – the non-quadratic functional J (due to Ψ) and the non-penetration conditions included in the definition of the convex set $\mathscr{K}$. Using the semismooth Newton method, we shall approximate Ψ by a quadratic functional using the Taylor expansion

$$\Psi(\mathbf{u}) \approx \Psi(\mathbf{u}^k) + \mathbf{F}(\mathbf{u}^k)^T(\mathbf{u}-\mathbf{u}^k) + \frac{1}{2}(\mathbf{u}-\mathbf{u}^k)^T\mathsf{K}(\mathbf{u}^k)(\mathbf{u}-\mathbf{u}^k)$$

defined for a given approximation $\mathbf{u}^k \in \mathscr{K}$ of the solution $\mathbf{u}$ to the problem (17.3). Let us denote $\mathbf{f}_k = \mathbf{f} - \mathbf{F}(\mathbf{u}^k)$, $\mathsf{K}_k = \mathsf{K}(\mathbf{u}^k)$ and define:

$$\mathscr{K}_k = \mathscr{K} - \mathbf{u}^k = \left\{\mathbf{v} \in \mathbb{R}^n \ ; \ \mathsf{B}_E\mathbf{v} = \mathbf{o}, \ \mathsf{B}_I\mathbf{v} \leq \mathbf{c}_{I,k}, \ \mathbf{c}_{I,k} := \mathbf{c}_I - \mathsf{B}_I\mathbf{u}^k\right\},$$

$$J_k(\mathbf{v}) := \frac{1}{2}\mathbf{v}^T\mathsf{K}_k\mathbf{v} - \mathbf{f}_k^T\mathbf{v}, \quad \mathbf{v} \in \mathscr{K}_k. \tag{17.6}$$

Then, the Newton step reads

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \delta\mathbf{u}^k, \quad \mathbf{u}^{k+1} \in \mathscr{K},$$

where $\delta\mathbf{u}^k \in \mathscr{K}_k$ is a unique minimizer of J_k on $\mathscr{K}_k$,

$$J_k(\delta\mathbf{u}^k) \leq J_k(\mathbf{v}), \qquad \forall \mathbf{v} \in \mathscr{K}_k, \tag{17.7}$$

i.e., $\delta\mathbf{u}^k \in \mathscr{K}_k$ solves

$$\left(\mathsf{K}_k\delta\mathbf{u}^k\right)^T(\mathbf{v}-\delta\mathbf{u}^k) \geq \mathbf{f}_k^T(\mathbf{v}-\delta\mathbf{u}^k) \quad \forall \mathbf{v} \in \mathscr{K}_k. \tag{17.8}$$

Notice that if we substitute $\mathbf{v} = \mathbf{u}^{k+1} \in \mathscr{K}$ into (17.5), $\mathbf{v} = \mathbf{u} - \mathbf{u}^k \in \mathscr{K}_k$ into (17.8), and sum up, then we obtain the inequality

$$\left(\mathsf{K}(\mathbf{u}^k)\delta\mathbf{u}^k\right)^T(\mathbf{u}-\mathbf{u}^{k+1}) \geq \left(\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{u}^k)\right)^T(\mathbf{u}-\mathbf{u}^{k+1}),$$

which can be arranged into the form

$$(\mathbf{u}^{k+1} - \mathbf{u})^T \mathsf{K}(\mathbf{u}^k)(\mathbf{u}^{k+1} - \mathbf{u}) \leq \left(F(\mathbf{u}^k) - \mathbf{F}(\mathbf{u}) - \mathsf{K}(\mathbf{u}^k)(\mathbf{u}^k - \mathbf{u})\right)^T (\mathbf{u} - \mathbf{u}^{k+1}).$$

Hence, one can simply derive the local quadratic convergence of the semismooth Newton method provided that $\mathbf{u}^k$ is sufficiently close to $\mathbf{u}$.

17.3 Algorithms for Elasto-Plasticity

The discretized elasto-plastic problem can be solved by the following algorithm:

Algorithm 17.1 Solution of discretized elasto-plastic problem.

{*Initialization.*}
$\mathbf{u}_0 = \mathbf{o},\ \boldsymbol{\varepsilon}_{0,T} = \mathbf{o},\ \boldsymbol{\sigma}_{0,T} = \mathbf{o},\ \boldsymbol{\kappa}_{0,T} = \mathbf{o}$ for any $T \in \mathscr{T}_h$
{*Main loop.*}
for $k = 0, \ldots, n-1$
  find $\delta\mathbf{u}_{k+1} \in \mathscr{K}$: $\mathbf{F}_k(\delta\mathbf{u}_{k+1})^T(\mathbf{v} - \delta\mathbf{u}_{k+1}) = \delta\mathbf{f}_{k+1}^T(\mathbf{v} - \delta\mathbf{u}_{k+1}),\ \ \mathbf{v} \in \mathscr{K}$
  for $T \in \mathscr{T}_h$
    $\delta\boldsymbol{\varepsilon}_{k+1,T} = \mathsf{N}_T\mathsf{R}_T\delta\mathbf{u}_{k+1},\ \boldsymbol{\varepsilon}_{k+1,T} = \boldsymbol{\varepsilon}_{k,T} + \delta\boldsymbol{\varepsilon}_{k+1,T}$
    $\delta\boldsymbol{\sigma}_{k+1,T} = \mathbf{T}_{k,T}(\delta\boldsymbol{\varepsilon}_{k+1,T}),\ \boldsymbol{\sigma}_{k+1,T} = \boldsymbol{\sigma}_{k,T} + \delta\boldsymbol{\sigma}_{k+1,T}$
    $\delta\boldsymbol{\kappa}_{k+1,T} = \mathbf{T}_{\kappa,k,T}(\delta\boldsymbol{\varepsilon}_{k+1,T}),\ \boldsymbol{\kappa}_{k+1,T} = \boldsymbol{\kappa}_{k,T} + \delta\boldsymbol{\kappa}_{k+1,T}$
  end *for*
end *for*

We solve the nonlinear system of equations (17.5) by the semismooth Newton method (see, e.g., [5]). The corresponding algorithm reads as follows:

Algorithm 17.2 Semismooth Newton method.

{*Initialization.*}
$\delta\mathbf{u}_{k,0} = \mathbf{o}$ {*Main loop.*}
for $i = 0, 1, 2, \ldots$
  find $\delta\mathbf{u}_i \in \mathscr{K}_k$: $J_k(\delta\mathbf{u}^k) \leq J_k(\mathbf{v}) \qquad \forall \mathbf{v} \in \mathscr{K}_k$
  compute $\delta\mathbf{u}_{k,i+1} = \delta\mathbf{u}_{k,i} + \delta\mathbf{u}_i$
  if $\|\delta\mathbf{u}_{k,i+1} - \delta\mathbf{u}_{k,i}\| / (\|\delta\mathbf{u}_{k,i+1}\| + \|\delta\mathbf{u}_{k,i}\|) \leq \varepsilon_{\text{Newton}}$
    then *stop*
  end *if*
end *for*

Here $\varepsilon_{\text{Newton}} > 0$ is the relative stopping tolerance.

17.4 TFETI Method for Inner Problem and Benchmark

The inner problem (17.7) has the same structure as the primal frictionless TFETI problem (11.21). Thus, we can switch to dual as in Sect. 11.6 to get the dual problem to (17.7) in the form

$$\min \Theta(\boldsymbol{\lambda}) \quad \text{s.t.} \quad \boldsymbol{\lambda}_I \geq \mathbf{o} \quad \text{and} \quad \mathsf{R}_k^T(\mathbf{f} - \mathsf{B}^T\boldsymbol{\lambda}) = \mathbf{o}, \tag{17.9}$$

where

$$\Theta(\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\lambda}^T\mathsf{B}\mathsf{K}_k^+\mathsf{B}^T\boldsymbol{\lambda} - \boldsymbol{\lambda}^T(\mathsf{B}\mathsf{K}_k^+\mathbf{f}_k - \mathbf{c}_k), \tag{17.10}$$

R_k is a full rank matrix such that ImR_k=ImK_k, and

$$\mathsf{B} = \begin{bmatrix} \mathsf{B}_E \\ \mathsf{B}_I \end{bmatrix} \quad \mathbf{c} = \begin{bmatrix} \mathbf{c}_E \\ \mathbf{c}_{I,k} \end{bmatrix} = \begin{bmatrix} \mathbf{o} \\ \mathbf{c}_{I,k} \end{bmatrix}.$$

After the preconditioning and regularization of (17.10) by the TFETI projector P to the rigid body modes and regularization by $\rho\mathsf{Q}$, $\rho > 0$, and $\mathsf{Q} = \mathsf{I} - \mathsf{P}$ as in Sect. 11.8, we can solve the inner problem (17.7) in the same way as a contact problem with elastic bodies in Chap. 11, i.e., to use Algorithm 9.2 (SMALBE) to generate the iterations for the Lagrange multipliers for the equality constraints and to use Algorithm 8.2 (MPRGP) to solve the bound constrained problems in the inner loop. Since the stiffness matrices K_k, $k = 1, 2, \ldots$ are spectrally equivalent to the elastic matrices K that were introduced in Sect. 11.5 (see Čermák et al. [6] and Kienesberger, Langer, and Valdman [7]), it is possible to get for the solution of problem (17.7) by TFETI similar optimality results as those developed in Sect. 11.10.

Using the methodology described in Chap. 12, the algorithm can be modified to the solution of elasto-plastic contact problem with friction. Due to the arguments mentioned above, we can formulate for the solution of such problems with friction by TFETI similar optimality results as those developed in Sect. 12.9. Of course, the optimality results are valid only for the inner loop.

17.5 Numerical Experiments

We shall illustrate the performance of the algorithms presented in this chapter on the solution of a variant of the two-beams academic benchmark that was introduced in Sect. 11.11. The size of the beams is $3 \times 1 \times 1$ [m]. We use regular meshes generated in MatSol [8]. The Young modulus, the Poisson ratio, the initial yield stress for the von Mises criterion, and the hardening modulus are

$$E = 210{,}000\ [MPa], \quad \nu^i = 0.29, \quad \sigma_y^i = 450\ [MPa], \quad H_m^i = 10{,}000\ [MPa], \quad i = 1, 2.$$

The vertical traction 150 [MPa] was prescribed on the top face of the upper beam. The initial stress (or plastic strain) state was set to zero.

The problem discretized by 1,533,312/326,969 primal/dual variables and 384,000 finite elements was decomposed into 384 subdomains. The solution with 356,384 plastic elements required six Newton iterations, 67 SMALSE-M iterations, and 5,375 multiplications by the Hessian of the dual function. The precision of the Newton method and inner loop was controlled by

$$\frac{\|\mathbf{u}^{k+1} - \mathbf{u}^k\|}{\|\mathbf{u}^{k+1}\| + \|\mathbf{u}^k\|} \leq 10^{-4}$$

and by the relative precision of the inner loop 10^{-7}, respectively. The distribution of von Mises stress and total displacement for the finest mesh are in Figs. 17.1 and 17.2.

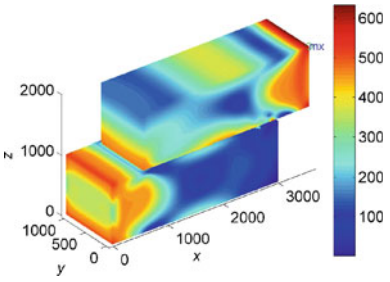

Fig. 17.1 von Mises stress distribution

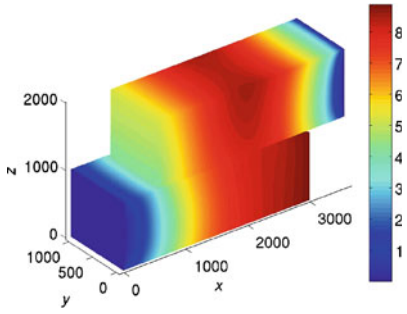

Fig. 17.2 Total displacement

17.6 Comments

The application of domain decomposition methods to the analysis of elasto-plastic bodies was considered, e.g., by Yagawa, Soneda, and Yoshimura [9] or Carstensen [10]. Here we followed Čermák [4] and Čermák and Sysala [3]; there can be found also the solution of an elasto-plastic variant of the yielding clamp connection (see Fig. 1.6). The adaptation of TFETI to the solution of the inner loop is straightforward and can be especially useful for the solution of large multibody problems. The proposed method can be used or can be a part of the solution of other contact inelastic problems, e.g., the analysis of von Mises' elasto-plastic bodies with isotropic hardening or loading with perfect plasticity [11, 12].

References

1. Han, W., Reddy, B.D.: Plasticity: Mathematical Theory and Numerical Analysis. Springer, New York (1999)
2. de Souza Neto, E.A., Perić, D., Owen, D.R.J.: Computational Methods for Plasticity, Theory and Applications. Wiley, West Sussex (2008)
3. Čermák, M., Sysala, S.: Total-FETI method for solving contact elasto-plastic problems. LNCSE **98**, 955–965 (2014)
4. Čermák, M.: Scalable algorithms for solving elasto-plastic problems. Ph.D. thesis, VŠB-TU Ostrava (2012)
5. Qi, L., Sun, J.: A nonsmooth version of Newton's method. Math. Program. **58**, 353–367 (1993)
6. Čermák, M., Kozubek, T., Sysala, S., Valdman, J.: A TFETI domain decomposition solver for elastoplastic problems. Appl. Math. Comput. **231**, 634–653 (2014)
7. Kienesberger, J., Langer, U., Valdman, J.: On a robust multigrid-preconditioned solver for incremental plasticity. In: Blaheta, R., Starý, J.(eds.) Proceedings of IMET 2004 – Iterative Methods, Preconditioning & Nummerical PDE, pp. 84–87. Institute of Geonics AS CR (2004)
8. Kozubek, T., Markopoulos, A., Brzobohatý, T., Kučera, R., Vondrák, V., Dostál, Z.: MatSol–MATLAB efficient solvers for problems in engineering (2015). http://industry.it4i.cz/en/products/matsol/
9. Yagawa, G., Soneda, N., Yoshimura, S.: A large scale finite element analysis using domain decomposition method on a parallel computer. Comput. Struct. **38**(5–6), 615–625 (1991)
10. Carstensen, C.: Domain decomposition for a non-smooth convex minimization problem and its application to plasticity. Numer. Linear Algebr. Appl. **4**(3), 177–190 (1997)
11. Čermák, M., Haslinger, J., Kozubek, T., Sysala, S.: Discretization and numerical realization of contact problems for elastic-perfectly plastic bodies. PART II - numerical realization, limit analysis. ZAMM -. J. Appl. Math. Mech. **95**(12), 1348–1371 (2015)
12. Sysala, S., Haslinger, J., Hlaváček, I., Čermák, M.: Discretization and numerical realization of contact problems for elastic-perfectly plastic bodies. PART I - discretization, limit analysis. ZAMM -. J. Appl. Math. Mech. **95**(4), 333–353 (2015)

Chapter 18
Contact Shape Optimization

18.1 Introduction

Contact shape optimization problems in 3D have a structure which can be effectively exploited by the TFETI-based methods introduced in Part III. The reason is that the preparation of the solution of the state problem can be reused in the solution of a number of auxiliary contact problems that arise in each design step.

Let us recall that the goal of the contact shape optimization is to find the shape of a part of the contact boundary that satisfies predefined constraints and minimizes a prescribed cost function. The unknown shape is defined by the *design variables* $\boldsymbol{\alpha} \in \mathscr{D}_{\text{admis}}$ defined for $\boldsymbol{\alpha}$ from an admissible set $\mathscr{D}_{\text{adm}} \subset \mathbb{R}^p$. The cost function $J\big(u(\boldsymbol{\alpha}), \boldsymbol{\alpha}\big)$ typically depends explicitly not only on the design variables $\boldsymbol{\alpha}$ but also on the displacement $u(\boldsymbol{\alpha})$, which corresponds to the solution of the associated contact problem. The latter problem is called the *state problem*. The state problem is a contact problem introduced in Part III, the specification of which depends on $\boldsymbol{\alpha}$. Thus the ith body of the state problem which corresponds to $\boldsymbol{\alpha}$ occupies in the reference configuration the domain $\Omega^i(\boldsymbol{\alpha})$, the corresponding boundary traction $\mathbf{f}_F^i(\boldsymbol{\alpha})$ can depend on $\boldsymbol{\alpha}$, etc. Denoting by $\mathscr{C}(\boldsymbol{\alpha})$ the set of the displacements $u(\boldsymbol{\alpha})$, which satisfy the conditions of equilibrium and possibly some additional constraints specified for the problem, the *shape optimization problem* reads

$$\text{Find } \overline{\boldsymbol{\alpha}} \in \mathscr{D}_{\text{adm}}, \quad J\big(u(\overline{\boldsymbol{\alpha}}), \overline{\boldsymbol{\alpha}}\big) \le J\big(u(\boldsymbol{\alpha}), \boldsymbol{\alpha}\big), \quad \boldsymbol{\alpha} \in \mathscr{D}_{\text{adm}},\ u(\boldsymbol{\alpha}) \in \mathscr{C}(\boldsymbol{\alpha}).$$

Here, we use TFETI to minimize the compliance of a system of bodies in mutual contact subject to box and volume constraints. The TFETI method is especially useful in the sensitivity analysis, the most expensive part of the design optimization process. We use the semi-analytical method that reduces the sensitivity analysis to a sequence of QP problems with the same Hessian matrix, so that the relatively expensive formulation of the dual problem can be reused (see, e.g., Dostál, Vondrák, and Rasmussen [1]) and many problems can be solved in parallel.

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3_18

18.2 Discretized Minimum Compliance Problem

Let us consider a variant of the Hertz 3D contact problems depicted in Fig. 18.1 and assume that the shape of the upper face of the lower body is controlled by the vector of design variables $\boldsymbol{\alpha} \in \mathscr{D}_{\text{adm}}$, $\mathscr{D}_{\text{adm}} = [0, 1]^p \subset \mathbb{R}^p$. Our goal is to change the shape of the lower body in order to minimize the compliance while preserving the volume of the bodies. The zero normal displacements are prescribed on the faces which share their boundaries with the axes.

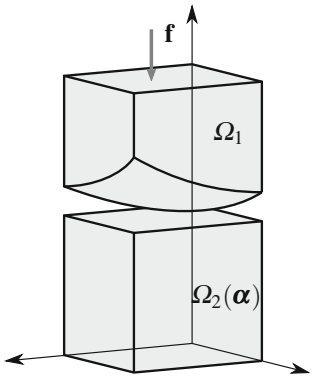

Fig. 18.1 Basic notation of the contact problem

After choosing the design vector $\boldsymbol{\alpha}$, we can decompose and discretize the problem in Fig. 18.1 as in Chap. 11. We get the energy functional in the form

$$J(\mathbf{u}, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{u}^T \mathsf{K}(\boldsymbol{\alpha})\mathbf{u} - \mathbf{u}^T \mathbf{f}(\boldsymbol{\alpha}), \tag{18.1}$$

where the stiffness matrix $\mathsf{K}(\boldsymbol{\alpha})$ and possibly the vector of external nodal forces $\mathbf{f}(\boldsymbol{\alpha})$ depend on $\boldsymbol{\alpha}$.

The matrix B_I and the vector $\mathbf{c}_I$ that describe the linearized condition of non-interpenetration also depend on $\boldsymbol{\alpha}$, so that the solution $\mathbf{u}(\boldsymbol{\alpha})$ of the state problem with the bodies that occupy the domains Ω_1 and $\Omega_2 = \Omega_2(\boldsymbol{\alpha})$ can be found as the solution of the minimization problem

$$\min J(\mathbf{u}, \boldsymbol{\alpha}) \quad \text{subject to} \quad \mathbf{u} \in \mathscr{C}_h(\boldsymbol{\alpha}), \tag{18.2}$$

where

$$\mathscr{C}_h(\boldsymbol{\alpha}) = \{\mathbf{u} : \mathsf{B}_I(\boldsymbol{\alpha})\mathbf{u} \leq \mathbf{c}_I(\boldsymbol{\alpha}) \text{ and } \mathsf{B}_E\mathbf{u} = \mathbf{o}\}.$$

Denoting the solution of (18.2) by $\mathbf{u}(\boldsymbol{\alpha})$, we can define a function

$$\mathscr{J}(\boldsymbol{\alpha}) = J(\mathbf{u}(\boldsymbol{\alpha}), \boldsymbol{\alpha})$$

to get the contact shape optimization problem to find

$$\min_{\boldsymbol{\alpha}\in\mathscr{D}_{\mathrm{adm}}} \mathscr{J}(\boldsymbol{\alpha}) = \min_{\boldsymbol{\alpha}\in\mathscr{D}_{\mathrm{adm}}} \min_{\mathbf{u}\in\mathscr{C}_h(\boldsymbol{\alpha})} J(\mathbf{u},\boldsymbol{\alpha}). \tag{18.3}$$

The set of admissible design variables $\mathscr{D}_{\mathrm{adm}}$ defines all possible designs. It has been proved that the minimal compliance problem has at least one solution and that the function $\mathscr{J}(\boldsymbol{\alpha})$ is differentiable with respect to $\boldsymbol{\alpha}$ under natural assumptions (see, e.g., Haslinger and Neittanmäki [2]).

18.3 Sensitivity Analysis

The goal of the sensitivity analysis is to find the influence of the design change on the solution of the state problem. The minimization of the cost function is then carried out by improving the initial design in a feasible decrease direction using $\nabla\mathbf{u}(\boldsymbol{\alpha})$ the columns of which are defined by the directional derivatives of the solutions of the state problem

$$\mathbf{u}'(\boldsymbol{\alpha},\mathbf{d}) = \lim_{t\to 0^+} \frac{\mathbf{u}(\boldsymbol{\alpha}+t\mathbf{d}) - \mathbf{u}(\boldsymbol{\alpha})}{t},$$

where for $\mathbf{d}$ are substituted the columns of the identity matrix I.

Here we shall briefly describe the semi-analytical sensitivity analysis. Some comparisons of the efficiency of alternative methods on practical problems in the context of domain decomposition can be found in Dostál, Vondrák, and Rasmussen [1] and Vondrák et al. [3].

Let us recall the Lagrange function of the TFETI state problem (18.2) has the form

$$L(\mathbf{u},\boldsymbol{\lambda},\boldsymbol{\alpha}) = \frac{1}{2}\mathbf{u}^T\mathsf{K}(\boldsymbol{\alpha})\mathbf{u} - \mathbf{f}^T(\boldsymbol{\alpha})\mathbf{u} + \boldsymbol{\lambda}^T\big(\mathsf{B}(\boldsymbol{\alpha})\mathbf{u} - \mathbf{c}(\boldsymbol{\alpha})\big), \tag{18.4}$$

where

$$\mathsf{B}(\boldsymbol{\alpha}) = \begin{bmatrix} \mathsf{B}_I(\boldsymbol{\alpha}) \\ \mathsf{B}_E^T \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1(\boldsymbol{\alpha}) \\ \vdots \\ \mathbf{b}_m(\boldsymbol{\alpha}) \end{bmatrix}$$

and $\mathbf{u}$ and $\boldsymbol{\lambda}$ depend on the vector of design variables $\boldsymbol{\alpha}$, so that the KKT conditions for (18.2) read

$$\mathsf{K}(\boldsymbol{\alpha})\mathbf{u} - \mathbf{f}(\boldsymbol{\alpha}) + \mathsf{B}^T(\boldsymbol{\alpha})\boldsymbol{\lambda} = \mathbf{o} \quad \mathsf{B}_E\mathbf{u} = \mathbf{o}, \tag{18.5}$$

$$\mathsf{B}_I(\boldsymbol{\alpha})\mathbf{u} - \mathbf{c}_I(\boldsymbol{\alpha}) \le \mathbf{o}, \quad \boldsymbol{\lambda}_I \ge \mathbf{o}. \tag{18.6}$$

To find the gradient of the cost function at the solution $\mathbf{u}$ of the state problem (18.2), let us first consider the derivative $\mathbf{u}'(\boldsymbol{\alpha},\mathbf{d})$ of $\mathbf{u} = \mathbf{u}(\boldsymbol{\alpha})$ in a given direction $\mathbf{d}$ at

$\mathbf{u} = \mathbf{u}(\boldsymbol{\alpha})$, let $\mathscr{I}$ and $\mathscr{E}$ denote the indices of the equality and inequality constrains, respectively, $\mathscr{I} \cup \mathscr{E} = \{1, \ldots, m\}$, let us denote by $\mathscr{A} = \mathscr{A}(\mathbf{u})$ the active set of $\mathbf{u}$,

$$\mathscr{A} = \left\{ i \in \{1, \ldots, m\} : \mathbf{b}_i \mathbf{u} = c_i \right\},$$

and let us decompose $\mathscr{A}$ into the weakly active set $\mathscr{A}_w$ and the strongly active set $\mathscr{A}_s$,

$$\mathscr{A}_w = \{ i \in \mathscr{I} : \lambda_i = 0 \}, \quad \mathscr{A}_s = \mathscr{A} \setminus \mathscr{A}_w,$$

where $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\alpha})$ denotes the vector of Lagrange multipliers associated with $\mathbf{u}$. After the differentiation of the left equation in the direction $\mathbf{d}$, we get that $\mathbf{z} = \mathbf{u}'(\boldsymbol{\alpha}, \mathbf{d})$ satisfies

$$\mathsf{K}(\boldsymbol{\alpha})\mathbf{z} - \mathbf{f}'(\boldsymbol{\alpha}, \mathbf{d}) + \mathsf{K}'(\boldsymbol{\alpha}, \mathbf{d})\mathbf{u} + \mathsf{B}'(\boldsymbol{\alpha}, \mathbf{d})^T \boldsymbol{\lambda} = \mathbf{o}, \tag{18.7}$$

so that $\mathbf{u}'(\boldsymbol{\alpha}, \mathbf{d})$ can be obtained by the solution of

$$\min \ \widetilde{J}_{\boldsymbol{\alpha},\mathbf{d}}(\mathbf{z}) \quad \text{subject to} \quad \mathbf{z} \in \mathscr{D}(\boldsymbol{\alpha}, \mathbf{d}), \tag{18.8}$$

where

$$\widetilde{J}_{\boldsymbol{\alpha},\mathbf{d}}(\mathbf{z}) = \frac{1}{2}\mathbf{z}^T \mathsf{K}(\boldsymbol{\alpha})\mathbf{z} - \mathbf{z}^T \left(\mathbf{f}'(\boldsymbol{\alpha}, \mathbf{d}) + \mathsf{K}'(\boldsymbol{\alpha}, \mathbf{d})\mathbf{u} + \mathsf{B}'(\boldsymbol{\alpha}, \mathbf{d})^T \boldsymbol{\lambda} \right)$$

and

$$\begin{aligned}
\mathscr{D}(\boldsymbol{\alpha}, \mathbf{d}) &= \mathscr{D}_w(\boldsymbol{\alpha}, \mathbf{d}) \cap \mathscr{D}_s(\boldsymbol{\alpha}, \mathbf{d}), \\
\mathscr{D}_w(\boldsymbol{\alpha}, \mathbf{d}) &= \{ \mathbf{z} \in \mathbb{R}^n : \mathbf{b}_i(\boldsymbol{\alpha})\mathbf{z} \le c_i'(\boldsymbol{\alpha}, \mathbf{d}) - \mathbf{b}_i'(\boldsymbol{\alpha}, \mathbf{d})\, \mathbf{u}(\boldsymbol{\alpha}) \text{ for } i \in \mathscr{A}_w \}. \\
\mathscr{D}_s(\boldsymbol{\alpha}, \mathbf{d}) &= \{ \mathbf{z} \in \mathbb{R}^n : \mathbf{b}_i(\boldsymbol{\alpha})\mathbf{z} = c_i'(\boldsymbol{\alpha}, \mathbf{d}) - \mathbf{b}_i'(\boldsymbol{\alpha}, \mathbf{d})\, \mathbf{u}(\boldsymbol{\alpha}) \ \text{ for } \ i \in \mathscr{A}_s \}.
\end{aligned}$$

Recall that $\mathsf{K}'(\boldsymbol{\alpha}, \mathbf{d})$, $f'(\boldsymbol{\alpha}, \mathbf{d})$, $\mathsf{B}'(\boldsymbol{\alpha}, \mathbf{d})$, and $\mathbf{c}'(\boldsymbol{\alpha}, \mathbf{d})$ are the directional derivatives in the direction $\mathbf{d}$ that can be simply evaluated. It has been proved (see, e.g., Haslinger and Mäkinen [4]) that the solution of (18.8) is the directional derivative $\mathsf{u}'(\alpha, \mathsf{d})$ of the solution of the state problem (18.2).

Let us now assume that $\boldsymbol{\alpha}$, $\mathbf{d}$ are fixed and denote

$$\begin{aligned}
\widetilde{\mathbf{f}} &= \widetilde{\mathbf{f}}(\boldsymbol{\alpha}, \mathbf{d}) = \mathbf{f}'(\boldsymbol{\alpha}, \mathbf{d}) - \mathsf{K}'(\boldsymbol{\alpha}, \mathbf{d})\mathbf{u} - \mathsf{B}'(\boldsymbol{\alpha}, \mathbf{d})^T \boldsymbol{\lambda}, \\
\mathbf{c}_w &= \mathbf{c}_w(\boldsymbol{\alpha}, \mathbf{d}) = \mathbf{c}'_{\mathscr{A}_w}(\boldsymbol{\alpha}, \mathbf{d}) - \mathsf{B}'_{\mathscr{A}_w}(\boldsymbol{\alpha}, \mathbf{d})\mathbf{u}, \\
\mathbf{c}_s &= \mathbf{c}_s(\boldsymbol{\alpha}, \mathbf{d}) = \mathbf{c}'_{\mathscr{A}_s}(\boldsymbol{\alpha}, \mathbf{d}) - \mathsf{B}'_{\mathscr{A}_s}(\boldsymbol{\alpha}, \mathbf{d})\mathbf{u}, \\
\mathsf{B}_w &= \mathsf{B}_w(\boldsymbol{\alpha}) = \mathsf{B}_{\mathscr{A}_w}(\boldsymbol{\alpha}), \\
\mathsf{B}_s &= \mathsf{B}_s(\boldsymbol{\alpha}) = \mathsf{B}_{\mathscr{A}_s}(\boldsymbol{\alpha}), \\
\widetilde{\mathsf{B}} &= \begin{bmatrix} \mathsf{B}_w \\ \mathsf{B}_s \end{bmatrix}, \quad \widetilde{\mathbf{c}} = \begin{bmatrix} \mathbf{c}_w \\ \mathbf{c}_s \end{bmatrix},
\end{aligned}$$

where B_w and B_s are the matrices formed by the rows of B corresponding to the weak and strong constraints, respectively, and similarly for the vectors $\mathbf{c}_w$ and $\mathbf{c}_s$. Using the duality as in Sect. 11.6, we can get the dual problem

$$\min \Phi(\boldsymbol{\mu}) \quad \text{subject to} \quad \boldsymbol{\mu}_w \geq \mathbf{o} \quad \text{and} \quad \mathsf{R}^T(\boldsymbol{\alpha})\widetilde{\mathsf{B}}^T\widetilde{\boldsymbol{\mu}} = \mathsf{R}^T(\boldsymbol{\alpha})\widetilde{\mathbf{f}}, \tag{18.9}$$

where

$$\widetilde{\boldsymbol{\mu}} = \begin{bmatrix} \boldsymbol{\mu}_w \\ \boldsymbol{\mu}_s \end{bmatrix}$$

are the multipliers enforcing the weak and strong constraints and

$$\Phi(\boldsymbol{\mu}) = \frac{1}{2}\boldsymbol{\mu}^T\widetilde{\mathsf{B}}\mathsf{K}^{\dagger}(\boldsymbol{\alpha})\widetilde{\mathsf{B}}^T\boldsymbol{\mu} - \boldsymbol{\mu}^T\left(\widetilde{\mathsf{B}}\mathsf{K}^{\dagger}(\boldsymbol{\alpha})\widetilde{\mathbf{f}}^T - \widetilde{\mathbf{c}}\right).$$

Finally, the derivative $\mathbf{u}'(\boldsymbol{\alpha}, \mathbf{d})$ can be obtained by the procedure described in Sect. 3.7.

Problem (18.9) with the bound and linear equality constraints is efficiently solvable by the combination of the SMALBE and MPRGP algorithms. Notice that the semi-analytical method for the sensitivity analysis for one design step requires the solution of p quadratic programming problems (18.9) with the same matrix

$$\mathsf{K}^{+}(\boldsymbol{\alpha}) = \operatorname{diag}\left(\mathsf{K}_1^{+}(\boldsymbol{\alpha}), \ldots, \mathsf{K}_s^{+}(\boldsymbol{\alpha})\right).$$

Thus, we can reuse the factorization and the kernel of the matrix $\mathsf{K}(\boldsymbol{\alpha})$ from the solution of the state problem to the solution of the problems arising in the semi-analytical sensitivity analysis and one design step requires only one factorization of the stiffness matrix. Moreover, if the domains are reasonably discretized so that the assumptions of Theorem 11.2 are satisfied, then the sensitivity analysis of a design step has asymptotically linear complexity. Notice that the problems (18.9) for varying $\mathbf{d}_i \in \mathbb{R}^p$, $i = 1, \ldots, p$ are independent so that they can be solved in parallel.

18.4 Numerical Experiments

We have tested the performance of the whole optimization procedure on the minimization of the compliance

$$\mathscr{J}(\boldsymbol{\alpha}) = \frac{1}{2}\mathbf{u}(\boldsymbol{\alpha})^T\mathsf{K}(\boldsymbol{\alpha})\mathbf{u}(\boldsymbol{\alpha}) - \mathbf{u}(\boldsymbol{\alpha})^T\mathbf{f}(\boldsymbol{\alpha}) \tag{18.10}$$

of the Hertz system defined above subject to the upper bound on the volume of the lower body. The compliance was controlled by the shape of the top side of the lower body that is parameterized by means of $3 \times 3 = 9$ design variables defined over this face. The design variables were allowed to move in vertical directions within the

prescribed box-constraints. The results of semi-analytical sensitivity analysis were used in the inner loop of the sequential quadratic programming algorithm (SQP) (see e.g., Nocedal and Wright [5]). The distribution of the normal contact pressure of the initial and optimized system of bodies are in Figs. 18.2 and 18.3.

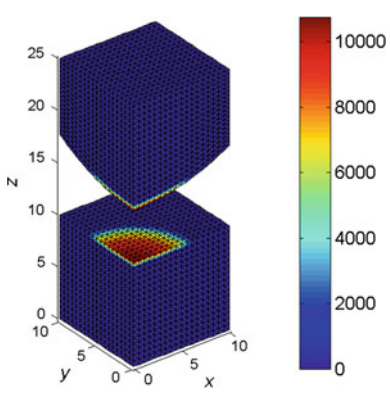

Fig. 18.2 Initial design – contact traction

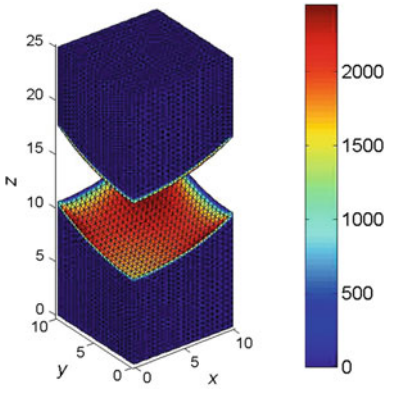

Fig. 18.3 Final design – contact traction

The computations were carried out with varying discretization and decomposition parameters h and H, respectively. The results concerning the solution of the state problem are in Table 18.1. The relative precision of the computations was 10^{-6}.

Table 18.1 State problem – performance for varying decomposition and discretization

Number of subdomains	2	16	54	128
Primal variables	24,576	196,608	663,552	1,572,864
Dual variables	11,536	23,628	92,232	233,304
Null space	12	96	324	768
Hessian mults.	64	153	231	276
Outer iterations	12	16	13	10

The optimized shape was observed after 121 design steps. The history of the decrease of the cost function is in Fig. 18.4. More details about this problem can be found in Vondrák et al. [3].

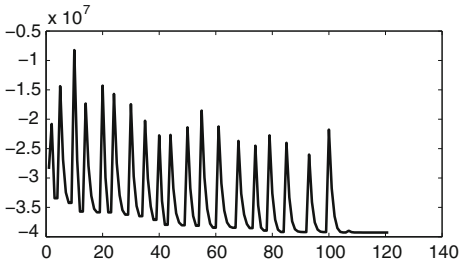

Fig. 18.4 3D Hertz – minimization history

18.5 Comments

The contact shape optimization problems have been studied by several authors and many results are known. The approximation theory and the existence results related to this problem can be found in Haslinger and Neittanmäki [4] or Haslinger and Mäkinen [4]. These authors also present basic semi-analytical sensitivity analysis for the problems governed by a variational inequality, describe its implementation based on a variant of the SQP (Sequential Quadratic Programming), and give numerical results for 2D problems. More engineering approach to constrained design sensitivity analysis is discussed also by Haug, Choi, and Komkov [6], for enhancing plasticity and friction see Kim, Choi, and Chen [7]. Younsi et al. [8], who obtained a numerical solution of 3D problem, exploited the relation between the discretized and continuous problem by combining the multilevel approach with genetic algorithms. Their approach can use parallelization on the highest level but does not fully exploit the structure of the problem. The natural two-level structure of the contact shape optimization problem for minimization of weight was exploited by Herskovits et al. [9]. The solution of 3D contact shape optimization problems with friction can be found in Beremlijski et al. [10].

The presentation of this chapter is based on Dostál, Vondrák, and Rasmussen [1] and Vondrák et al. [3]. Similar results as those presented here can be obtained with TBETI. Such approach would simplify remeshing as BETI uses surface grids on the boundary of the (sub)domains only.

References

1. Dostál, Z., Vondrák, V., Rasmussen, J.: FETI based semianalytic sensitivity analysis in contact shape optimization. In: Fast Solution of Discretized Optimization Problems. International Series of Numerical Mathematics, Basel, pp. 98–106 (2001)
2. Haslinger, J., Neittanmäki, P.: Finite Element Approximation for Optimal Shape, Material and Topology Design, 2nd edn. Wiley, New York (1996)
3. Vondrák, V., Kozubek, T., Markopoulos, A., Dostál, Z.: Parallel solution of 3D contact shape optimization problems based on Total FETI domain decomposition method. Struct. Multidiscip.

Optim. **42**(6), 955–964 (2010)
4. Haslinger, J., Mäkinen, R.A.E.: Introduction to Shape Optimization. SIAM, Philadelphia (2002)
5. Nocedal, J., Wright, S.F.: Numerical Optimization. Springer, New York (2000)
6. Haug, E.J., Choi, K., Komkov, V.: Design Sensity Analysis of Structural Systems. Academic Press, New York (1986)
7. Kim, N.H., Choi, K., Chen, J.S.: Shape design sensitivity analysis and optimization of elasto-plasticity with frictional contact. AIAA J. **38**, 1742–1753 (2000)
8. Younsi, R., Knopf-Lenoir, C., Selman, A.: Multi-mesh and adaptivity in 3D shape optimization. Comput. Struct. **61**(6), 1125–1133 (1996)
9. Herskovits, J., et al.: Contact shape optimization: a bilevel programming approach. Struct. Multidiscip. Optim. **20**(3), 214–221 (2000)
10. Beremlijski, P., Haslinger, J., Kočvara, M., Kučera, R., Outrata, J.: Shape optimization in three-dimensional contact problems with coulomb friction. SIAM J. Optim. **20**(1), 416–444 (2009)

Chapter 19
Massively Parallel Implementation

The FETI domain decomposition methods that are presented in this book can effectively exploit the parallel facilities provided by modern computers to the solution of very large problems, currently up to hundereds of billions of nodal variables provided they are implemented properly. However, this task is far from trivial and straightforward.

The FETI methods appeared in the early 90s, when the parallel computers were not assumed to have some tens or even hundreds of thousands of cores, and an immediate goal was to use them for the solution of the problems discretized by a few millions of the degrees of freedom. Thus it is not surprising that we face new problems. For example, the cost of the assembling of the projector to the "natural coarse grid," which is nearly negligible for smaller problems, starts to essentially affect the cost of the solution when the dimension of the dual problem reaches some tens of millions. New challenges are posed also by the emerging exascale technologies, the effective exploitation of which has to take into account a hierarchical organization of memory, the varying cost of operations depending on the position of arguments in memory, and the increasing role of communication costs. Last but not least, it is important to exploit an up-to-date software, either open source or commercial, as the effective implementation of some standard steps, such as the application of direct solvers, is highly nontrivial and affects the overall performance of algorithms.

Here, we present some hints concerning the parallel implementation of FETI-type algorithms for the solution of very large problems, including the implementation of the action of a generalized inverse K^+ of the stiffness matrix K and the action of the projector to the "natural coarse grid" P. We briefly discuss the possibility to overcome the bottleneck by introducing the third-level grid by a variant of HTFETI (Hybrid TFETI). The third level is introduced by the decomposition of TFETI subdomains into smaller subdomains that are partly glued in corners or by averages at the primal level (see e.g., Klawonn and Rheinbach [1]). We also briefly describe the packages that were used for the solution of the benchmarks throughout the book, namely MatSol based on parallel MATLAB, PERMON based on PETSc, and ESPRESO based on Intel MKL and Cilk.

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3_19

19.1 Stiffness Matrix Factorization and Action of K^+

The cost of TFETI iterations, either of the CG steps or of the gradient projection step, is typically dominated by the multiplication of a vector $\mathbf{x}$ by S^+, the generalized inverse of the Schur complement of the stiffness matrix K with respect to the indices of variables associated with the interior of the subdomains. Though there are attempts to implement this step by iterative solvers, it seems that such approach has a little chance to succeed due to the necessity to carry out this step in high precision. The same is true for the application of the projector P to the natural coarse grid, which requires the solution of the coarse problem with high precision.

The implementation of $\mathsf{K}^+\mathbf{x}$ can be parallelized without any data transfers because of the block-diagonal structure of K,

$$\mathsf{K} = \mathrm{diag}(\mathsf{K}_1, \ldots, \mathsf{K}_s),$$

so that it is easy to achieve that each CPU core deals with a local stiffness matrix K_i. If this is not possible, e.g., when the number of the subdomains is greater than the number of cores, then it is possible to allocate more blocks to one core. Unfortunately, the projector P does not possess the block-diagonal structure that is easy to implement in parallel. We shall discuss this topic in the next section.

A natural way how to exploit effectively the massively parallel computers is to maximize the number of subdomains so that the sizes of the subdomain stiffness matrices are reduced, which accelerates both their factorization and subsequent forward/backward solves. On the other hand, the negative effect of that is the increase of the dual and the coarse space dimensions. The coarse problem becomes a bottleneck when its dimension, which is equivalent to the dimension of $\mathrm{Ker}\mathsf{K}$, attains some ten thousands.

The performance of the algorithms is affected also by the choice of LU direct solver. The effect of the choice of the direct solver on the performance of TFETI in PERMON library (experiments with PETSc, MUMPS, and SuperLU on Cray XE6 machine HECToR) can be found in Hapla, Horák, and Merta [2]. To evaluate $\mathsf{K}^+\mathbf{x}$, each core regularizes a subdomain stiffness matrix using fixing nodes and then factorizes. The application of K^+ then consists of purely local backward and forward substitutions once per each TFETI iteration.

19.2 Coarse Problem Implementation – Action of P

The action time and the level of communication depend primarily on the implementation of the solution $\mathsf{G}\mathsf{G}^T\mathbf{x} = \mathbf{y}$. Here, we consider two strategies of the solution of coarse problems, namely using the LU factorization and applying the explicit inverse of $\mathsf{G}\mathsf{G}^T$. The explicit orthonormalization of G, which was conveniently used in the description of the algorithms, turned out to be less efficient.

19.2.1 Assembling GG^T in Parallel

The implementation of the action of the projector P to the coarse grid starts by assembling GG^T. If the number of subdomains is large, then it can be effectively generated in parallel using the block structure of R, B, in particular

$$\mathsf{R} = \mathrm{diag}(\mathsf{R}_1, \ldots, \mathsf{R}_s) \quad \text{and} \quad \mathsf{B} = [\mathsf{B}_1, \ldots, \mathsf{B}_s].$$

The ith core first obtains $\mathsf{R}_i, \mathsf{B}_i$ and generates $\mathsf{G}_i = \mathsf{R}_i^T \mathsf{B}_i^T$ of G. Then the ith core receives G_j from the neighboring subdomains, generates in parallel $\mathsf{G}_i \mathsf{G}_j^T = [\mathsf{GG}^T]_{ij}$, and sends the result to the master node, where GG^T is assembled and prepared for the factorization. Since GG^T is symmetric, it follows that for each couple of neighboring subdomains, only the subdomain with a smaller index gets the corresponding part of G from the subdomain with a greater index j.

19.2.2 Parallel Explicit Inverse

The approach used in the ESPRESO library is based on hybrid parallelization. It implements the observation that the factorization of GG^T is significantly less time consuming than executing the forward and backward substitution m-times, where m is the order of GG^T. The assembled GG^T is broadcasted from the master compute node (the node with MPI rank equal to 0) to all compute nodes. The factorization is then carried out on each compute node using the threaded sparse direct solver. During this step the identical workload is executed on every node and does not cause any slowdown. Modern sparse direct solvers also contain the threaded implementation of the solve routine for multiple right-hand sides. This combination leads to the efficient utilization of all computational resources. Moreover, since each node computes only that part of the inverse of GG^T that is needed by that node, there is no additional communication.

19.2.3 Parallel Direct Solution

An alternative approach chosen in the PERMON library splits the N_c cores of the global communicator into the groups of cores called subcommunicators in MPI. The number of the subcommunicators is N_r (number of cores doing redundant work), so that the number of cores in each subcommunicator is equal to N_c/N_r. The GG^T matrix can be factorized sequentially on a master core or in parallel using, e.g., MUMPS or SuperLU_DIST in subcommunicators. The matrix–vector multiplication $(\mathsf{GG}^T)^{-1}\mathbf{x}$ then consists of the backward and forward substitutions, which are not local and a considerable amount of communication is needed in each iteration, but

the parallelization reduces the memory requirements – there are nearly no memory limits as more and more cores can be engaged in the subcommunicators. The optimal number of cores per subcommunicator for our problems is $\sqrt[3]{N_c^2}$ based on numerical experiments.

19.3 Hybrid TFETI (HTFETI)

Even if there are several efficient parallelization strategies, there is still a limit on the dimension of the coarse problem. It has been observed that we can get beyond this limit if we aggregate a small number of neighboring subdomains into the clusters, which naturally results in the smaller coarse problem. The aggregation of the subdomains into the clusters can be enforced partly by the Lagrange multipliers and partly by identifying the corner nodes as in Fig. 19.1 and/or enforcing zero face and/or edge averages on the primal level. The result is the HTFETI method (see, e.g., Klawonn and Rheinbach [1], Říha et al. [3]). The subdomains in clusters are typically joined in such a way that the clusters have the same dimension of rigid body modes as each of their floating subdomains (notice that the dimension of the rigid modes of the cluster in Fig. 19.1 is the same as that of any of its floating subdomains). The HTFETI decomposition is characterized by the decomposition parameters $H_c > H_s$ which correspond to the diameters of the clusters and their subdomains, respectively.

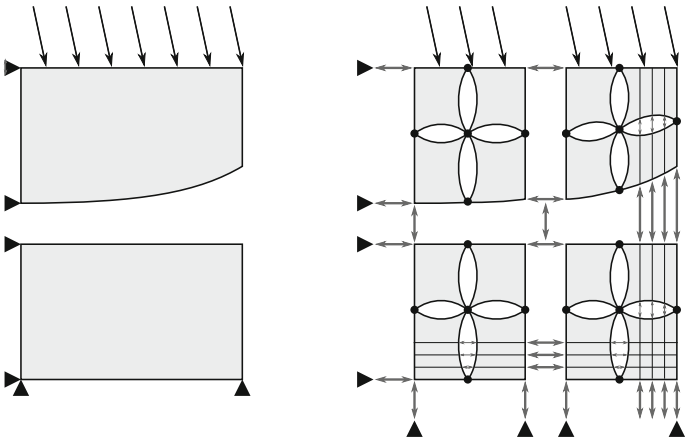

Fig. 19.1 Cluster of subdomains joined in corners

This approach allows the parallelization of the original problem into up to hundreds of thousands of cores, which is not reachable with the standard FETI methods. Thus it enables to exploit the whole current petascale supercomputers. It is not difficult to see that the optimality results which we have proved for the TFETI methods

can also be proved for the HTFETI methods provided H_c/H_s and H_s/h is uniformly bounded, where h is the parameter of the discretization of the subdomains. However, the convergence of the HTFETI method is a little slower as compared with the TFETI method, especially when no edge or face averages are enforced on the primal level, so that its application is worth the complications only for the solution of very large problems.

19.3.1 Description of TFETI Method

Let us illustrate the HTFETI method on the analysis of a 2D cantilever beam decomposed into four subdomains and two clusters as in Fig. 19.2.

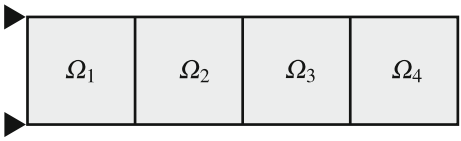

Fig. 19.2 2D cantilever beam decomposed into four subdomains

After the discretization and domain decomposition, the resulting linear system reads

$$\left[\begin{array}{cccc|cc|c} \mathsf{K}_1 & \mathsf{O} & \mathsf{O} & \mathsf{O} & \mathsf{B}_{c,1}^T & \mathsf{O} & \mathsf{B}_1^T \\ \mathsf{O} & \mathsf{K}_2 & \mathsf{O} & \mathsf{O} & \mathsf{B}_{c,2}^T & \mathsf{O} & \mathsf{B}_2^T \\ \mathsf{O} & \mathsf{O} & \mathsf{K}_3 & \mathsf{O} & \mathsf{O} & \mathsf{B}_{c,3}^T & \mathsf{B}_3^T \\ \mathsf{O} & \mathsf{O} & \mathsf{O} & \mathsf{K}_4 & \mathsf{O} & \mathsf{B}_{c,4}^T & \mathsf{B}_4^T \\ \hline \mathsf{B}_{c,1} & \mathsf{B}_{c,2} & \mathsf{O} & \mathsf{O} & \mathsf{O} & \mathsf{O} & \mathsf{O} \\ \mathsf{O} & \mathsf{O} & \mathsf{B}_{c,3} & \mathsf{B}_{c,4} & \mathsf{O} & \mathsf{O} & \mathsf{O} \\ \hline \mathsf{B}_1 & \mathsf{B}_2 & \mathsf{B}_3 & \mathsf{B}_4 & \mathsf{O} & \mathsf{O} & \mathsf{O} \end{array}\right] \left[\begin{array}{c} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \\ \mathbf{u}_4 \\ \hline \boldsymbol{\lambda}_{c,1} \\ \boldsymbol{\lambda}_{c,2} \\ \hline \boldsymbol{\lambda} \end{array}\right] = \left[\begin{array}{c} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \mathbf{f}_3 \\ \mathbf{f}_4 \\ \hline \mathbf{o} \\ \mathbf{o} \\ \hline \mathbf{c} \end{array}\right], \tag{19.1}$$

where B_i and $\mathsf{B}_{c,i}$, $i = 1, \ldots, 4$, denote the blocks of the standard gluing constraints and the copies of their rows which correspond to the components of $\boldsymbol{\lambda}$ acting on the corners associated with Ω_i, respectively. Thus if the redundant rows of

$$\mathsf{B}_c = \left[\begin{array}{cccc} \mathsf{B}_{c,1} & \mathsf{B}_{c,2} & \mathsf{O} & \mathsf{O} \\ \mathsf{O} & \mathsf{O} & \mathsf{B}_{c,3} & \mathsf{B}_{c,4} \end{array}\right]$$

are omitted, the primal solution components remain the same. Let us permute (19.1) to get

$$\left[\begin{array}{ccc|ccc|c} \mathsf{K}_1 & \mathsf{O} & \mathsf{B}_{c,1}^T & \mathsf{O} & \mathsf{O} & \mathsf{O} & \mathsf{B}_1^T \\ \mathsf{O} & \mathsf{K}_2 & \mathsf{B}_{c,2}^T & \mathsf{O} & \mathsf{O} & \mathsf{O} & \mathsf{B}_2^T \\ \mathsf{B}_{c,1} & \mathsf{B}_{c,2} & \mathsf{O} & \mathsf{O} & \mathsf{O} & \mathsf{O} & \mathsf{O} \\ \hline \mathsf{O} & \mathsf{O} & \mathsf{O} & \mathsf{K}_3 & \mathsf{O} & \mathsf{B}_{c,3}^T & \mathsf{B}_3^T \\ \mathsf{O} & \mathsf{O} & \mathsf{O} & \mathsf{O} & \mathsf{K}_4 & \mathsf{B}_{c,4}^T & \mathsf{B}_4^T \\ \mathsf{O} & \mathsf{O} & \mathsf{O} & \mathsf{B}_{c,3} & \mathsf{B}_{c,4} & \mathsf{O} & \mathsf{O} \\ \hline \mathsf{B}_1 & \mathsf{B}_2 & \mathsf{O} & \mathsf{B}_3 & \mathsf{B}_4 & \mathsf{O} & \mathsf{O} \end{array}\right] \left[\begin{array}{c} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \boldsymbol{\lambda}_{c,1} \\ \hline \mathbf{u}_3 \\ \mathbf{u}_4 \\ \boldsymbol{\lambda}_{c,2} \\ \hline \boldsymbol{\lambda} \end{array}\right] = \left[\begin{array}{c} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \mathbf{o} \\ \hline \mathbf{f}_3 \\ \mathbf{f}_4 \\ \mathbf{o} \\ \hline \mathbf{c} \end{array}\right]. \tag{19.2}$$

For convenience, let us rewrite (19.2) in the block form which corresponds to the line partition as

$$\left[\begin{array}{cc|c} \widetilde{\mathsf{K}}_1 & \mathsf{O} & \widetilde{\mathsf{B}}_1^T \\ \mathsf{O} & \widetilde{\mathsf{K}}_2 & \widetilde{\mathsf{B}}_2^T \\ \hline \widetilde{\mathsf{B}}_1 & \widetilde{\mathsf{B}}_2 & \mathsf{O} \end{array}\right] \left[\begin{array}{c} \widetilde{\mathbf{u}}_1 \\ \widetilde{\mathbf{u}}_2 \\ \hline \widetilde{\boldsymbol{\lambda}} \end{array}\right] = \left[\begin{array}{c} \widetilde{\mathbf{f}}_1 \\ \widetilde{\mathbf{f}}_2 \\ \hline \widetilde{\mathbf{c}} \end{array}\right]. \tag{19.3}$$

Eliminating $\widetilde{\mathbf{u}}_i$, $i = 1, 2$, we also eliminate the subset of dual variables $\boldsymbol{\lambda}_{c,j}$, $j = 1, 2$, related to the matrix B_c. Thus the structure looks like a problem decomposed into two clusters – the first and second subdomains belong to the first cluster, the third and fourth subdomains belong to the second cluster. The rigid body modes (rotations) associated with the subdomains and clusters are depicted in Fig. 19.3.

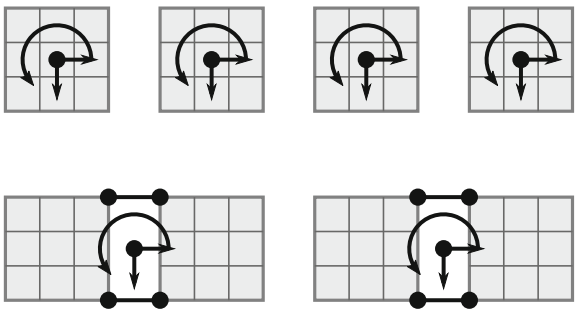

Fig. 19.3 Cluster of subdomains joined in corners

Matrices $\widetilde{\mathsf{K}}_1, \widetilde{\mathsf{K}}_2$ can be interpreted as the cluster stiffness matrices with the kernels $\widetilde{\mathsf{R}}_1, \widetilde{\mathsf{R}}_2$, respectively. Denoting

$$\begin{array}{lll} \widetilde{\mathsf{K}} = \mathrm{diag}(\widetilde{\mathsf{K}}_1, \widetilde{\mathsf{K}}_2), & \widetilde{\mathsf{B}} = [\widetilde{\mathsf{B}}_1, \widetilde{\mathsf{B}}_2], & \widetilde{\mathsf{R}} = \mathrm{diag}(\widetilde{\mathsf{R}}_1, \widetilde{\mathsf{R}}_2), \\ \widetilde{\mathsf{F}} = \widetilde{\mathsf{B}}\widetilde{\mathsf{K}}^{+}\widetilde{\mathsf{B}}^T, & \widetilde{\mathsf{G}} = \widetilde{\mathsf{R}}^T\widetilde{\mathsf{B}}^T, & \\ \widetilde{\mathbf{d}} = \widetilde{\mathsf{B}}\widetilde{\mathsf{K}}^{+}\widetilde{\mathbf{f}} - \widetilde{\mathbf{c}}, & \widetilde{\mathbf{e}} = \widetilde{\mathsf{R}}^T\widetilde{\mathbf{f}}, & \end{array}$$

we get the Schur complement system

$$\begin{bmatrix} \widetilde{\mathsf{F}} & -\widetilde{\mathsf{G}}^T \\ -\widetilde{\mathsf{G}} & \mathsf{O} \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{\lambda}} \\ \widetilde{\boldsymbol{\alpha}} \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{d}} \\ -\widetilde{\mathbf{e}} \end{bmatrix}, \tag{19.4}$$

which can be solved by the methods used by the classical FETI. The dimension of $\widetilde{\mathsf{G}}\widetilde{\mathsf{G}}^T$ is six, i.e., less than twelve corresponding to FETI.

The primal variables and corner multipliers are eliminated by the implicit factorization. We shall illustrate the procedure on the solution of the linear system $\widetilde{\mathsf{K}}_1\widetilde{\mathbf{x}}_1 = \widetilde{\mathbf{b}}_1$, i.e.,

$$\begin{bmatrix} \mathsf{K}_{1:2} & \mathsf{B}_{c,1:2}^T \\ \mathsf{B}_{c,1:2} & \mathsf{O} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{z} \end{bmatrix}, \tag{19.5}$$

where

$$\mathsf{K}_{1:2} = \mathrm{diag}(\mathsf{K}_1, \mathsf{K}_2), \qquad \mathsf{B}_{c,1:2} = [\mathsf{B}_{c,1}, \mathsf{B}_{c,2}].$$

Although (19.5) can be interpreted as a FETI problem, we solve it by elimination. The generalized Schur complement system reads as

$$\begin{bmatrix} \mathsf{F}_{c,1:2} & -\mathsf{G}_{c,1:2}^T \\ -\mathsf{G}_{c,1:2} & \mathsf{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \mathbf{d}_{c,1:2} \\ -\mathbf{e}_{c,1:2} \end{bmatrix}, \tag{19.6}$$

where

$$\mathsf{F}_{c,1:2} = \mathsf{B}_{c,1:2}\mathsf{K}_{1:2}^{+}\mathsf{B}_{c,1:2}^T, \quad \mathsf{G}_{c,1:2} = \mathsf{R}_{1:2}^T\mathsf{B}_{c,1:2}^T, \quad \mathbf{d}_{c,1:2} = \mathsf{B}_{c,1:2}\mathsf{K}_{1:2}^{+}\mathbf{b} - \mathbf{z},$$
$$\mathbf{e}_{c,1:2} = \mathsf{R}_{1:2}^T\mathbf{b}, \quad \mathsf{R}_{1:2} = \mathrm{diag}\,(\mathsf{R}_1, \mathsf{R}_2)\,.$$

To obtain $\widetilde{\mathbf{x}}_1$, both systems (19.5), (19.6) are solved in three steps:

$$\begin{aligned} \boldsymbol{\beta} &= \mathsf{S}_{c,1:2}^{+}\left(\mathbf{e}_{c,1:2} - \mathsf{G}_{c,1:2}\mathsf{F}_{c,1:2}^{-1}\mathbf{d}_{c,1:2}\right), \\ \boldsymbol{\mu} &= \mathsf{F}_{c,1:2}^{-1}\left(\mathbf{d}_{c,1:2} + \mathsf{G}_{c,1:2}^T\boldsymbol{\beta}\right), \\ \mathbf{x}_1 &= \mathsf{K}_{1:2}^{+}\left(\mathbf{b} - \mathsf{B}_{c,1:2}^T\boldsymbol{\mu}\right) + \mathsf{R}_{1:2}\boldsymbol{\beta}, \end{aligned} \tag{19.7}$$

where $\mathsf{S}_{c,1:2} = \mathsf{G}_{c,1:2}\mathsf{F}_{c,1:2}^{-1}\mathsf{G}_{c,1:2}^T$ is a singular Schur complement matrix.

The kernel $\widetilde{\mathsf{R}}_1$ of $\widetilde{\mathsf{K}}_1$ is the last object to be effectively evaluated. The orthogonality condition $\widetilde{\mathsf{K}}_1\widetilde{\mathsf{R}}_1 = \mathsf{O}$ can be written in the form

$$\begin{bmatrix} \mathsf{K}_{1:2} & \mathsf{B}_{c,1:2}^T \\ \mathsf{B}_{c,1:2} & \mathsf{O} \end{bmatrix} \begin{bmatrix} \mathsf{R}_{1:2} \\ \mathsf{O} \end{bmatrix} \mathsf{H}_{1:2} = \begin{bmatrix} \mathsf{O} \\ \mathsf{O} \end{bmatrix}, \quad \widetilde{\mathsf{R}}_1 = \begin{bmatrix} \mathsf{R}_{1:2} \\ \mathsf{O} \end{bmatrix} \mathsf{H}_{1:2}. \tag{19.8}$$

Assuming that the subdomain kernels R_1 and R_2 are known, it remains to determine $\mathsf{H}_{1:2}$. However, the first equation in (19.8) does not impose any condition onto $\mathsf{H}_{1:2}$ and the second equation gives

$$\mathsf{B}_{c,1:2}\mathsf{R}_{1:2}\mathsf{H}_{1:2} = \mathsf{G}_{c,1:2}^T\mathsf{H}_{1:2} = \mathsf{O}. \tag{19.9}$$

It follows that $\mathsf{H}_{1:2}$ is the kernel of $\mathsf{G}_{c,1:2}^T$, which is not a full-column rank matrix due to the absence of Dirichlet boundary conditions in $\mathsf{B}_{c,1:2}$. To get matrix $\mathsf{H}_{1:2}$ efficiently, temporary sparse matrix $\mathsf{G}_{c,1:2}\mathsf{G}_{c,1:2}^T$ is assembled and factorized by the routine for singular matrices. Such routine provides not only the factors but also the kernel of $\mathsf{G}_{c,1:2}\mathsf{G}_{c,1:2}^T$, in this case the matrix $\mathsf{H}_{1:2}$. For more details see [4].

Preprocessing in HTFETI starts in the same way as in FETI by preparing the factors of K_i by direct solver and kernels R_i for each subdomain. Then one pair consisting of $\mathsf{F}_{c,j:k}$ and $\mathsf{S}_{c,j:k}$ is assembled and factorized on each cluster by a direct solver. The dimension of $\mathsf{F}_{c,1:2}$ is controlled by the number of Lagrange multipliers $\lambda_{c,1}$ gluing the cluster subdomains. The dimension of $\mathsf{S}_{c,1:2}$ is given by the sum of the defects of all matrices K_i belonging to the particular cluster.

19.3.2 Parallel Implementation

In our implementation, the HTFETI decomposition is mapped to the hardware in the following way. The clusters are mapped to the compute nodes, therefore the parallelization model for the distributed memory is used. In our case we use the message passing model (MPI). The subdomains inside the cluster are mapped to the CPU cores of the particular compute node; therefore the shared memory model is used for the second level. Our implementation allows us to have multiple clusters associated with a single compute node, but a single cluster cannot be processed on more than one node, as this is the limitation of the shared memory parallelization.

There are two major parts of the HTFETI solver that affect its parallel performance and scalability, the communication layer and the inter-cluster processing routines. The communication level deals with the optimization of the TFETI algorithm to minimize its communication overhead caused mainly by the gluing matrix $\widetilde{\mathsf{B}}$ multiplication and the application of the projector P. To have a fully optimized communication layer is essential for both TFETI and HTFETI methods.

19.3.3 Numerical Experiment

In order to eliminate the effect of nonlinearity of contact conditions, let us illustrate the power of HTFETI on the analysis of elastic cube without contact computed on Titan, a supercomputer at the Oak Ridge National Laboratory. It has 18,688 nodes each containing a 16-core AMD Opteron 6274 CPU with 32 GB of DDR3 memory

and an Nvidia Tesla K20X GPU with 6 GB GDDR5 memory. There are a total of 299,008 processor cores, and a total of 693.6 TiB of CPU and GPU RAM. The results with iterative solver accelerated by the lumped preconditioner are depicted in two graphs.

A weak scalability test is depicted in Fig. 19.4 including all relevant information. Number of degrees of freedom ranges from 30 millions to 70 billions and number of nodes from 8 to 17, 576.

The strong scalability test is shown on Fig. 19.5. The undecomposed problem was discretized by 11 billion nodal variables.

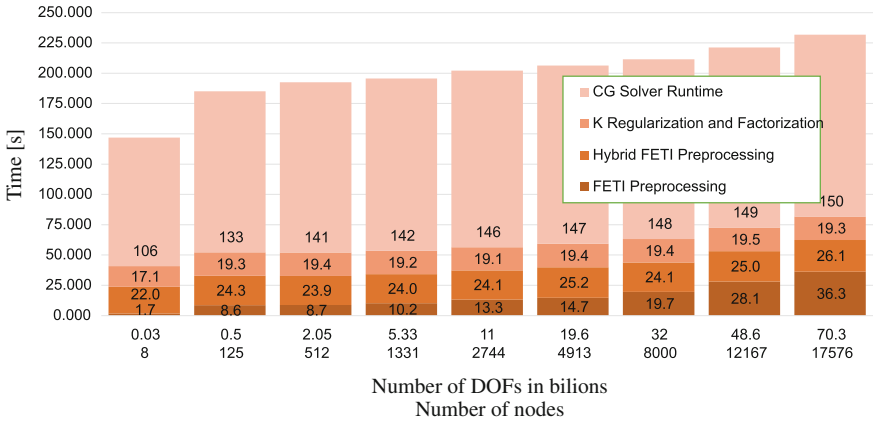

Fig. 19.4 HTFETI—linear elasticity—weak scalability test

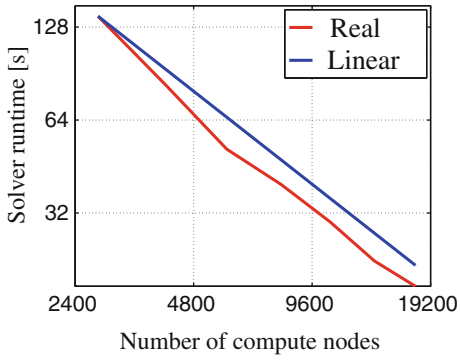

Fig. 19.5 HTFETI - linear elasticity—11 billion DOFs—strong scalability test

19.4 Communication Layer Optimization

The HTFETI solver can be implemented using the hybrid parallelization, which is well suited for multi-socket and multicore compute nodes provided by the hardware architecture in most of today's supercomputers. The first level of parallelization is designed for the parallel processing of the clusters of subdomains. Each cluster must be assigned to a single node, but if necessary, multiple clusters can be processed per node. The distributed memory parallelization is done using MPI. The essential part of this parallelization is the communication layer. This layer is identical whether the solver runs in the TFETI or HTFETI mode and comprises:

1. the parallel coarse problem solution using the distributed inverse matrix which merges two global communication operations (Gather and Scatter) into one (All-Gather),
2. the optimized version of global gluing matrix multiplication (matrix B for TFETI and $\bar{\mathsf{B}}_1$ for HTFETI)—written as a stencil communication which is fully scalable.

The stencil communication for simple decomposition into four subdomains is shown in Figs. 19.6 and 19.7, where the Lagrange multipliers that connect different neighboring subdomains are. In each iteration, whenever the multipliers are updated, the exchange is performed between the neighboring subdomains to finish the update. This affinity also controls the distribution of the data for the main distributed iterative solver, which iterates over the local multipliers only. In our implementation, each MPI process modifies only those elements of the vectors used by the solver that match the multipliers associated with the particular domain in the case of TFETI or the set of domains in a cluster in the case of HTFETI. We call this operation the vector compression. In the preprocessing stage, the local part of the gluing matrix is also compressed using the same approach (in this case it means that the empty rows are removed from the matrix), so that we can directly use the sparse matrix vector multiplication on the compressed vectors.

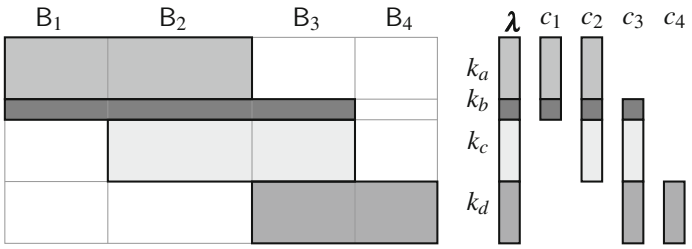

Fig. 19.6 Stencil communication: distribution of B and $\boldsymbol{\lambda}$ on cores c_i

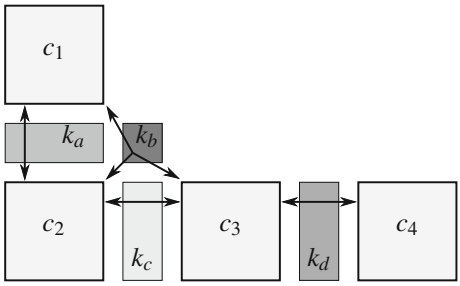

Fig. 19.7 Stencil communication: necessary data exchange k_i between cores

19.4.1 TFETI Hybrid Parallelization

In pure MPI implementation, each process holds the data of one subdomain, i.e., the number of cores equals the number of subdomains. The structure of the primal data and their distribution is shown in Fig. 19.8 left. The matrices K and R have a nice block-diagonal structure. The diagonal blocks are represented by sequential matrix types.

The $\mathsf{K}^{+}\mathbf{y}$ action occurs in each TFETI iteration and is typically implemented with a direct sparse solver for the sake of robustness. The factorization of the regularized K is performed in the setup phase. All these computations do not require any MPI communication.

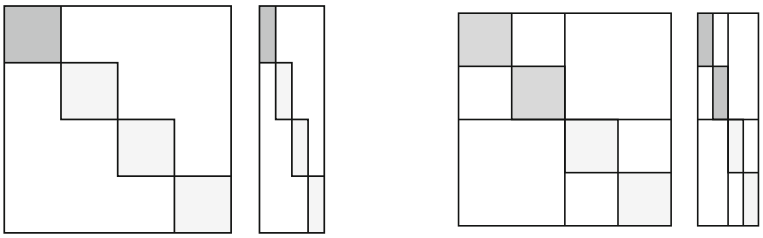

Fig. 19.8 Parallel distribution of K and R matrices. One core—one subdomain (*left*) and one core—more subdomains (*right*)

For the hybrid parallelization of TFETI with threads, it is necessary to add a new matrix type holding the array of diagonal blocks. Matrices K, K^{+}, and R can then be represented in a hybrid way by distributed matrices the diagonal blocks of which are sequential matrices. It is possible to hold an arbitrary number of subdomains on one MPI process with this extension. The domain can be divided into any number of subdomains regardless the number of available cores. The splitting of the original subdomain into smaller subdomains has a positive effect on the condition number of the dual operator so the number of iterations decreases. Moreover, the dimensions of

matrices K_i are smaller so their factorizations are faster at the cost of the increasing size of the coarse problem.

As mentioned in the previous section, three approaches to solve the coarse problem are used. The approach with the direct solver was also used and tested for the hybrid implementation. The coarse problem can be again solved fully parallel, using either a smaller number of processes or a single process. Thus, a direct solver using the shared memory parallelism can be used on a single process or the hybrid parallelism on more processes. The hybrid parallelization of TFETI enables solving the coarse problem on a single node in parallel and without communication. It also reduces MPI overhead if the coarse problem is solved on more processes.

19.5 MatSol, PERMON, and ESPRESO Libraries

19.5.1 MatSol

The algorithms described in Part II and III were developed, implemented, and tested in MatSol library. The Matsol library serves as a referential implementation of these algorithms and was the fundamental tool in our research. To parallelize the algorithms, MATLAB Distributed Computing Server and Parallel Computing Toolbox are used. Hence, the MatSol has full functionality to solve efficiently large problems of mechanics.

The solution process starts from the model which is either in the model database or it is converted to the model database from the standard commercial or noncommercial preprocessors such as ANSA, ANSYS, COMSOL, etc. The list of preprocessing tools is not limited and additional tools can be plugged into the library by creating a proper database convertor.

The preprocessing part proceeds depending on the problem solved. User can solve the static or transient analysis, the optimization problems, the problems in linear and nonlinear elasticity, and contact problems. For discretization, the finite or boundary element methods are used. As the domain decomposition techniques, the TFETI and TBETI methods are implemented. The decomposition into subdomains is done using Metis and spectral methods.

The solution process could be run either in the sequential or parallel mode. The solution algorithms are implemented in such a way that the code is the same for both sequential and parallel mode. MatSol library also includes the tools for postprocessing and postplotting. The results are then converted through the model database into the modeling tools for further postprocessing.

19.5.2 PERMON

PERMON (Parallel, Efficient, Robust, Modular, Object-oriented, Numerical) is a software package which aims at the massively parallel solution of problems of constrained quadratic programming (QP). PERMON is based on PETSc and combines aforementioned TFETI method and QP algorithms. The core solver layer consists of the PermonQP package for QP and its PermonFLLOP extension for FETI. PermonQP supports the separation of QP problems, their transformations, and solvers. It contains all QP solvers described in this book.

More can be found on PERMON website: https://permon.it4i.cz.

19.5.3 ESPRESO

ESPRESO is an ExaScale PaRallel FETI SOlver developed at IT4Innovations. The aim is to create a highly efficient parallel solver. Apart from the algorithms used by MatSol and Permon, it also enhances the HFETI method, which is designed to run on massively parallel machines with thousands of compute nodes and hundreds of thousands of CPU cores. The algorithms can be seen as a multilevel FETI method designed to overcome the main bottleneck of standard FETI methods, a large coarse problem, which arise when solving large problems decomposed into the large number of subdomains. ESPRESO can exploit modern many-core accelerators.

There are three major versions of the solver. ESPRESO CPU is a CPU version that uses the sparse representation of system matrices. It contains an efficient communication layer on the top of MPI 3.0 combined with the shared memory parallelization inside nodes. The communication layer was developed specifically for FETI solvers and uses several state-of-the-art communication hiding and avoiding techniques to achieve better scalability.

The ESPRESO solver can take advantage of many-core accelerators to speedup the solver runtime. To achieve this, it uses a dense representation of sparse system matrices in the form of Schur complements. The main advantage of using this approach in FETI solvers is the reduction of the iteration time. Instead of calling a solve routine of the sparse direct solver in every iteration, which by its nature is a sequential operation, the solver can use the dense matrix–vector multiplication (GEMV) routine. The GEMV offers the parallelism required by many-core accelerators and delivers up to $4\times$ speedup depending on the hardware configuration. There are two versions: ESPRESO MIC for Intel Xeon Phi and ESPRESO GPU for graphic accelerators.

More information can be found on ESPRESO website: https://espreso.it4i.cz.

19.6 Numerical Experiments

For the demonstration of weak parallel scalability of the algorithms presented in Part II and III, we used their implementation in the ESPRESSO software (see Sect. 19.5.3). As a benchmark, we consider a frictionless rigid punch problem depicted in Fig. 19.9 left. The rigid punch is pressed against the elastic cube $1 \times 1 \times 1$ [m] which is fixed at the bottom. The punch, which is defined in a reference configuration by the function

$$z(x_1, x_2) = 1.11 - 0.1 * \sin(\pi * x_1),$$

drops vertically by 0.11 [m]. The material properties are defined by Young's modulus $E = 2.1 \times 10^5$ [MPa] and Poisson's ratio $\mu = 0.3$. The the deformed cube is in Fig. 19.9 right.

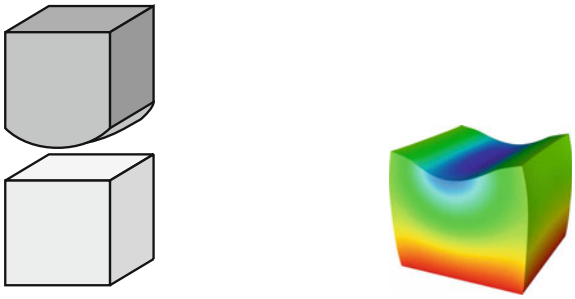

Fig. 19.9 Frictionless rigid punch problem (*left*) and deformed cube (*right*)

All computations were done using the Salomon cluster, operated by the National Supercomputer Center in Ostrava. The code was compiled with Intel C++ Compiler 15.0.2, and linked against Intel MPI Library for Linux 4.1 and PETSc 3.5.3. Salomon consists of 1,008 compute nodes, totaling 24,192 compute cores with 129TB RAM and giving over 2 Pflop/s theoretical peak performance. Each node is a powerful x86-64 computer, equipped with 24 cores with at least 128 GB RAM. The nodes are interconnected by the 7D Enhanced hypercube Infiniband network and equipped with the Intel Xeon E5-2680v3 processors. The Salomon cluster consists of 576 nodes without accelerators and 432 nodes equipped with the Intel Xeon Phi MIC accelerators.

The computations were carried out with varying decompositions into 64, 1,728, 8,000, 21,952, and 64,000 subdomains. When it was necessary, several subdomains were assigned to a single core. The stopping criterion was defined by the relative precision of the projected gradient and the feasibility error equal to 10^{-4} (measured in the Euclidean norm, see (9.40)) and compared with the Euclidean norm of the dual linear term. The results of computations are summarized in Table 19.1.

Table 19.1 Results of the benchmark—cube contact linear elasticity problem

N_S	n [10^6]	Hessian mult.	Iter. time [s]
64	5,719,872	73	258
1,728	154,436,544	83	311
8,000	714,984,000	117	364
21,952	1,961,916,096	142	418
64,000	5,719,872 000	149	497

The results vizualized in Fig. 19.10 illustrate good numerical and parallel scalability.

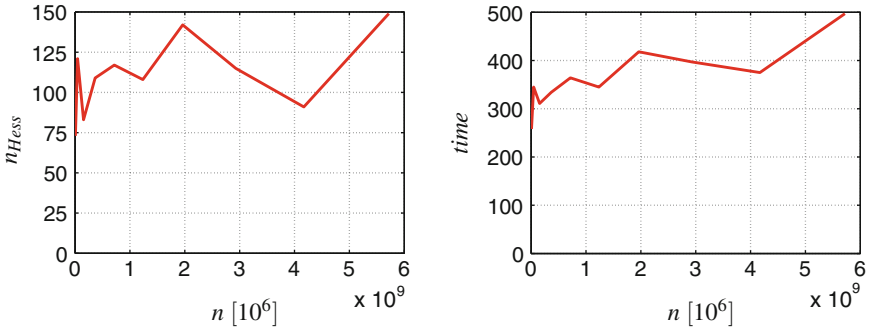

Fig. 19.10 Cube benchmark. Numbers of matrix–vector multiplications by F (*left*) and solution times (*right*)

References

1. Klawonn, A., Rheinbach, O.: Highly scalable parallel domain decomposition methods with an application to biomechanics. ZAMM - J. Appl. Math. Mech. **90**(1), 5–32 (2010)
2. Hapla, V., Horák, D., Merta, M.: Use of direct solvers in TFETI massively parallel implementation. Lect. Notes Comput. Sci. **14**, 192–205 (2013)
3. Říha, L., Brzobohatý, T., Markopoulos, A., Meca, O., Kozubek, T.: Massively parallel hybrid total FETI (HTFETI) solver. In: Proceedings of the Platform for Advanced Scientific Computing Conference - PASC 16, Article No. 7
4. Markopoulos, A., Říha, L., Brzobohatý, T., Jirůtková, P., Kučera, R., Meca, O., Kozubek, T.: Treatment of singular matrices in the Hybrid total FETI method. Lecture Notes in Computational Science and Engineering (to appear)

Bibliography

1. MUMPS Web page. http://graal.ens-lyon.fr/MUMPS/
2. SuperLU Web page. http://acts.nersc.gov/superlu/

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3

Index

Z. Dostál et al., *Scalable Algorithms for Contact Problems*, Advances in Mechanics and Mathematics 36, DOI 10.1007/978-1-4939-6834-3