Springer Series in Computational Mathematics 47

Sören Bartels

Numerical Methods for Nonlinear Partial Differential Equations

Springer

Springer Series in Computational Mathematics

Volume 47

More information about this series at http://www.springer.com/series/797

Sören Bartels

Numerical Methods for Nonlinear Partial Differential Equations

Springer

Sören Bartels
Abteilung für Angewandte Mathematik
Albert-Ludwigs-Universität Freiburg
Freiburg
Germany

ISSN 0179-3632
Springer Series in Computational Mathematics
ISBN 978-3-319-13796-4 ISBN 978-3-319-13797-1 (eBook)
DOI 10.1007/978-3-319-13797-1

Library of Congress Control Number: 2014957868

Mathematics Subject Classification (2010): 65N30, 65N12, 65M12, 65M22, 65K15, 49M15, 49M29

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Contents

Part I Analytical and Numerical Foundations

Part II Approximation of Classical Formulations

Part III Methods for Extended Formulations

Chapter 1
Introduction

> *As Henri Poincaré once remarked, "solution of a mathematical problem" is a phrase of indefinite meaning. Pure mathematicians sometimes are satisfied with showing that the nonexistence of a solution implies a logical contradiction, while engineers might consider a numerical result as the only reasonable goal. Such one-sided views seem to reflect human limitations rather than objective values.*
>
> – Richard Courant, 1941

1.1 Differential Equations and Numerical Methods

Starting with the brachistochrone problem solved by Johann Bernoulli in the 17th century, differential equations have become an indispensable tool to model, understand, and solve real world problems. The general approach via the Euler–Lagrange equations related to a variational principle and Euler's method to approximately solve differential equations discovered in the 18th century have provided a methodology that is still the basis for the mathematical modeling and numerical solution of many problems describing the behavior of solids and fluids. An important part of those models is the Laplace equation which is the Euler–Lagrange equation for the Dirichlet energy. The validity of the Dirichlet principle that postulates the existence of minimizers and hence solutions of Euler–Lagrange equations led to a controversial but constructive discussion in the 19th century. Important observations and objections from Cauchy, Riemann, Weierstraß, Schwarz, Ritz, and many others resulted in the birth of the calculus of variations and the finite element method. These mathematical concepts were led to concise mathematical theories by Hilbert and Courant at the beginning of the 20th century.

The abstract investigation of function spaces with topologies, functionals, and linear operators in the 20th century associated with deep theorems due to Banach and others, led to formulating the direct method in the calculus of variations in an

S. Bartels, *Numerical Methods for Nonlinear Partial Differential Equations*,
Springer Series in Computational Mathematics 47,
DOI 10.1007/978-3-319-13797-1_1

abstract way. This enabled the study of challenging mathematical problems involving the efficient description of processes on nanoscales and to establishing rigorous connections between classical models like three-dimensional elasticity and special applications involving the deformation of lower-dimensional objects. The development of functional analysis also had an impact in understanding numerical methods. A milestone is the construction of compatible finite element spaces and the rigorous analysis of problems involving linear constraints or higher-order derivatives within the inf-sup condition as an application of the closed range theorem.

The development of mixed and adaptive finite element methods, multigrid and domain decomposition algorithms, as well as wavelet and other compression techniques, in combination with the rapidly increasing available computer power in the last 50 years, have made it possible to compute forces acting on a bridge, turbulences created by the flow of air around an airplane, or compression of digital objects within minutes or seconds on personal computers. In parallel, the calculus of variations has become a powerful mathematical discipline that provides a framework for the effective description of complicated phenomena, such as microstructures in crystalline solids or the formation of cracks in materials with abstract mathematical objects such as measures. The price of these impressive individual advances of initially closely linked disciplines has led to a separation from them. Many powerful numerical techniques are difficult to analyze, while a lot of abstract analytical concepts are hard to realize practically. This book aims at contributing towards closing this gap.

1.2 Guidelines for the Development of Approximation Schemes

The classical numerical analysis of approximation schemes for partial differential equations exploits the concepts of stability and regularity. These are strong requirements that are available for certain classes of linear elliptic and parabolic equations. In many modern applications, solutions may be neither unique nor regular, and stability has to be formulated in a weaker sense. Possible concepts are weak convergence methods and convergence of minimizers. These approaches avoid making unrealistic regularity or uniqueness assumptions but justify rigorous approximation schemes. Since this leads to asymptotic statements, the efficiency and practical accuracy of these schemes then needs to be studied separately. We formulate some guidelines to justify a numerical discretization scheme.

The discretization of a variational problem or partial differential equation called a continuous problem typically leads to a family of finite-dimensional minimization problems:

$$\text{Minimize } I_h(u_h) \text{ in the set of functions } u_h \in \mathscr{A}_h$$

or equations

$$\text{Find } u_h \in \mathscr{A}_h \text{ such that } F_h(u_h) = 0$$

parametrized by a mesh-size $h > 0$. The main tasks in the development and justification of discretizations are the following:
(a) *Qualitative accuracy*: A numerical solution u_h should capture the relevant features of an exact solution u with a small number of degrees of freedom, e.g., if ε is a characteristic length scale of the problem under consideration, then the numerical method should provide qualitatively correct solutions with a computational length scale $h \approx \varepsilon/10$.
(b) *Efficient solution*: There should exist an iterative and convergent method that approximately solves the discrete formulation with a computational effort that scales like a low-order polynomial in the number of degrees of freedom, e.g., a fixed-point method or a gradient flow converges within a finite number of iterations and provides an approximation of a discrete solution with a solution error comparable to a power of h.
(c) *Asymptotic convergence*: A relevant quantity related to approximate solutions should converge to a corresponding quantity for the continuous problem as $h \to 0$, e.g., the approximate minimal energies $I_h(u_h)$ converge to $\inf_{u \in \mathscr{A}} I(u)$, subsequences of approximations converge to solutions of the continuous problem, or certain quantities $\sigma_h = \Sigma_h(u_h)$ converge to a meaningful quantity σ as $h \to 0$.

We investigate these requirements for certain prototypical nonlinear partial differential equations arising in the mathematical modeling of contact, phase transitions, ferromagnetism, bending, microstructures, fracture, and plasticity. The related model problems have in common that the regularity or uniqueness of solutions fails in general, so that numerical schemes have to be carefully developed in order to meet the aforementioned criteria. Short MATLAB implementations are included for most of the investigated algorithms that allow for testing the dependence of the performance on discretization parameters and for experimentally determining the typical preasymptotic range of the methods to thereby judge their qualitative accuracy.

1.3 Analytical and Numerical Foundations

The contents of this monograph are divided into three parts. The first provides the analytical framework for the considered model problems, collects the basic results related to the finite element method, and formulates abstract concepts for the analysis and solution of discretized problems. In the second part, the numerical solution of classical nonlinear partial differential equations, such as problems with inequality constraints, singularly perturbed parabolic equations, variational formulations with smooth constraints, and problems involving higher-order derivatives are discussed. In the third part, the approximation of solutions of nonstandard variational models is discussed on the basis of nonconvex minimization problems, extended formulations

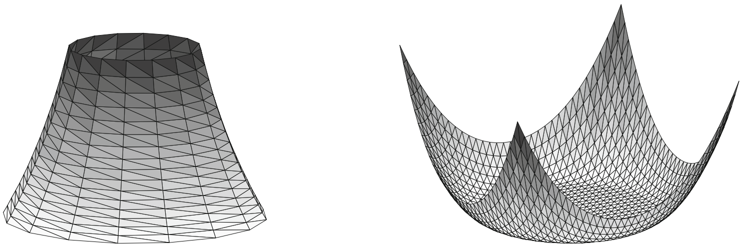


Fig. 1.1 Discrete minimal surface computed with a descent method (*left*) and solution of an obstacle problem computed with a semismooth Newton iteration (*right*); the plots were produced with the MATLAB routines `min_surf.m` and `obstacle_newton.m`

allowing for solutions with strong discontinuities, and nonsmooth, rate-independent evolution problems.

The mathematical model problems studied in this book are introduced and briefly described in the first chapter. All of them lead to minimization problems defined in certain function spaces or to evolution problems that are gradient flows of functionals. The basic techniques to study the existence and uniqueness of solutions can be addressed in the direct method in the calculus of variations and the concept of gradient or subdifferential flows that are described in Chap. 2. Chapter 3 introduces the finite element method and its analysis for linear elliptic and parabolic problems. Various auxiliary estimates such as inverse inequalities and density results as well as error estimates in different norms are recalled. In Chap. 4 general abstract methods for analyzing discretized problems and their iterative solutions are formulated. Key concepts are the variational convergence of discrete minimization problems to a corresponding continuous one, and the convergence of discretized partial differential equations. The different performance of iterative solution methods is illustrated by computing discrete minimal surfaces with large, unbounded gradients, as shown in the left plot of Fig. 1.1.

1.4 Approximation of Classical Formulations

The obstacle problem is a classical mathematical model problem that serves to understand inequality constraints in partial differential equations. The existence and uniqueness of solutions, the justification of numerical schemes with a priori and a posteriori error estimates, and the iterative solution with semismooth Newton methods are discussed in Chap. 5. The right plot of Fig. 1.1 shows the numerical solution of a two-dimensional obstacle problem with circular contact zone.

The evolution of an interface separating two phases of a substance is often based on the introduction of a phase field variable. The corresponding dynamics are modeled by the gradient flow of an energy functional with nonconvex lower-order terms leading to semilinear parabolic partial differential equations involving a critical para-

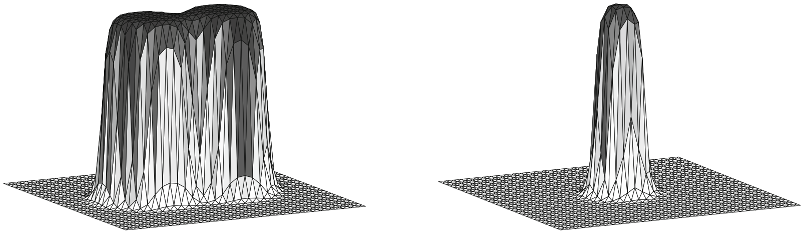


Fig. 1.2 Snapshots of a phase field variable in an Allen–Cahn evolution computed with a semi-implicit approximation scheme; the numerical solution was obtained with the MATLAB program `ac_linearized_euler.m`

meter that determines the width of the interfaces. Standard discretization methods provide useful approximations but error estimates typically depend exponentially on the inverse of the critical parameter. Chapter 6 provides an approach that leads to robust error estimates. When the critical parameter tends to zero, a so-called sharp interface model can be identified that determines the evolution of the interface in terms of its geometric properties. Figure 1.2 shows snapshots of an evolution for different times. The initial interface deforms into a sphere whose radius decreases gradually and eventually collapses. This is a simplified model for the description of certain melting processes.

The character of a problem changes significantly when a pointwise equality constraint is imposed, e.g., when a vector field attains its values in the zero level set of a given function. This leads to the notion of harmonic maps which are discussed in Chap. 7. Since neither uniqueness nor global regularity results are available, only the accumulation of approximations with respect to a weak topology at exact solutions can be shown. The pointwise constraint is imposed at the nodes of a triangulation and various iterative methods that either preserve the constraint via properties of the underlying partial differential equation or approximate it by a linearization and a subsequent optional projection are discussed. Figure 1.3 displays views of a harmonic map from a two-dimensional square into the two-dimensional unit sphere.

A dimension reduction from three-dimensional hyperelasticity in the bending regime leads to the nonlinear Kirchhoff model for the description of large deforma-

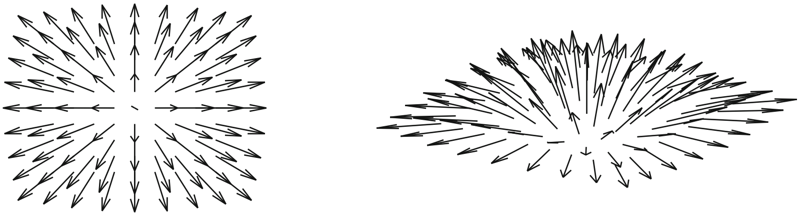


Fig. 1.3 Discrete harmonic map into the unit sphere viewed from two different perspectives; the iterative scheme realized in the MATLAB program `h1_flow_hm.m` led to the plots

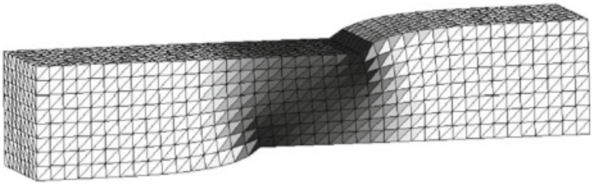


Fig. 1.4 Concentration of plastic strains in an elastoplastic compression experiment with small hardening parameter; the numerical solution of the nonlinear, nonsmooth evolution problem was computed with the MATLAB routine `elastoplasticity.m`

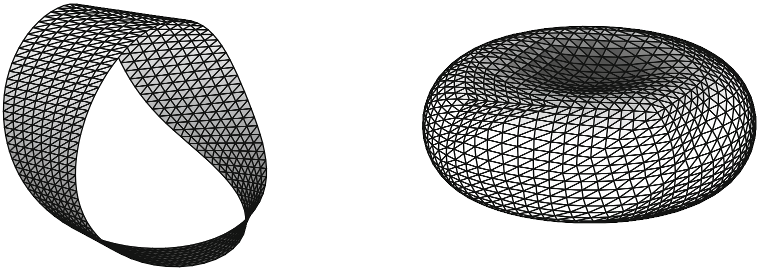


Fig. 1.5 Isometric deformation of a flat rectangular strip (*left*) and stationary configuration of the bending energy among closed surfaces with prescribed area and enclosed volume (*right*); the solutions were computed with the MATLAB routines `kirchhoff_nonlinear.m` and `willmore_helfrich_flow.m`

tions of two-dimensional objects, e.g., the bending of a sheet of paper. In the reduced model, local angle and area relations are preserved by the deformation and this is mathematically described by a pointwise isometry constraint which can be treated with techniques developed for harmonic maps. The additional numerical difficulty that higher-order derivatives have to be treated can be solved by employing finite element methods that were originally developed for linear bending problems describing small displacements. The resulting numerical scheme is provided in Chap. 8 and can be employed to compute the Möbius strip as the deformation of a flat strip of minimal bending energy subject to Dirichlet type boundary conditions at the ends of the strip. The result of a simulation with only a few degrees of freedom is shown in Fig. 1.5. A closely related problem consists in minimizing the Willmore energy in closed surfaces of a prescribed surface area and enclosed volume. Starting a gradient flow with an oblate spheroid, the gradient flow becomes stationary at a discocyte configuration that resembles the shape of a red blood cell. A low-order finite element scheme produced the plot shown in Fig. 1.5.

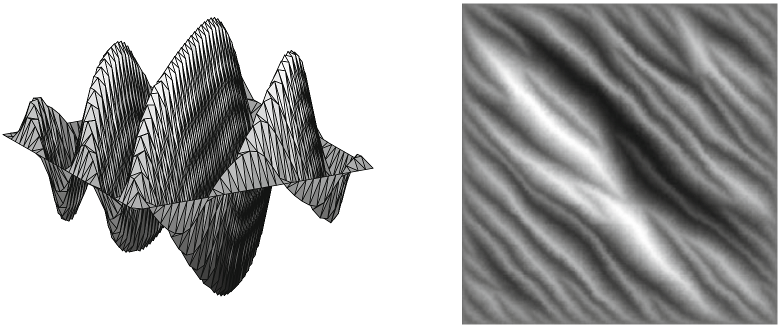


Fig. 1.6 Direct numerical minimization of a nonconvex energy functional leading to mesh-dependent oscillations on a coarse (*left*) and on a fine grid (*right*); the steepest descent method that computed the local minimizers is realized in the MATLAB program `energy_minimization.m`

1.5 Numerical Methods for Extended Formulations

The third part of the monograph investigates problems that are in general ill-posed within the classical framework of Sobolev spaces, e.g., provided by the direct method in the calculus of variations when an energy functional is coercive and weakly lower semicontinuous, and the underlying space is reflexive. Simple mathematical models for crystalline phase transitions lead to nonconvex energy densities which violate the semicontinuity requirement. Infimizing sequences still exist, but their accumulation points are in general no minimizers. This phenomenon is related to the occurrence of oscillations in the infimizing sequences that compensate for the nonconvexity of the energy functional. The weak limits and statistical information about the oscillations, described by Young measures, are relevant quantities in applications and their numerical approximation is discussed in Chap. 9. Figure 1.6 shows the results of a direct numerical approach based on minimizing the nonconvex energy functional with a descent method. The numerical approximations develop oscillations whose frequencies increase when the mesh is refined. The obtained configurations cannot be expected to be global minimizers, making their practical relevance unclear. Therefore, other approaches are required to obtain meaningful information.

Another reason for the failure of the direct method in the calculus of variations is the nonreflexivity of the employed space. Applications in image processing and and fracture mechanics motivate considering energy densities with linear growth and extended function spaces need to be considered that allow for minimizers with strong discontinuities. The appropriate discretization, error estimates, and iterative solution methods for certain model problems are studied in Chap. 10. Figure 1.7 shows as an application the denoising of an image obtained by minimizing an energy that preserves the edges of the noisy image.

The focus of Chap. 11 is on nonsmooth evolution problems occurring in the mathematical modeling of elastoplastic material behavior. Viscous regularizations of evolution problems are of limited practical use in applications. Instead, approxi-

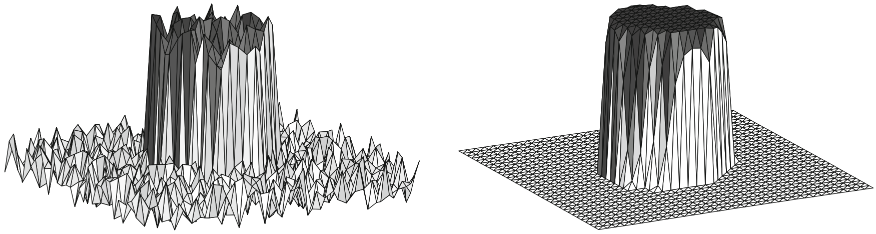


Fig. 1.7 Regularization of a noisy image (*left*) that preserves the edges of the image that are related to large, maximal gradients in the numerical approximation (*right*); the configurations were obtained with an iterative scheme realized in the MATLAB code `tv_reg.m`

mation schemes have to be developed which preserve the relevant property of rate-independence. Error estimates for implicit discretizations of the evolution problem in the case of kinematic and isotropic hardening are analyzed and their practical realization is discussed. The result of a numerical simulation based on the provided MATLAB codes displayed in Fig. 1.4 shows the tendency towards the development of a slip band in a compression experiment with small hardening parameter.

1.6 Objectives and Acknowledgments

The purpose of this work is to enable advanced students and experienced researchers to gain an entry into the numerical analysis of problems related to modern applications in continuum mechanics. Every chapter in the second and third part of the book starts with a review of the analytical properties of the model problem under consideration and gives selected references to provide links to detailed explanations and proofs. The main emphasis is on the development and analysis of approximation schemes that meet the general criteria outlined above. MATLAB implementations of the schemes that allow for a treatment of two- and three-dimensional problems are typically discussed at the end of the chapters. The codes are available at http://extras.springer.com/2015/978-3-319-13796-4. These implementations are meant to illustrate the simplicity and efficiency of the proposed schemes and to serve as reference codes that are easily accessible. Only a selected number of references is listed at the end of every chapter that serve as entries into the large number of relevant contributions.

The presented material is the result of several years of research, inspiration from numerous researchers' work, and discussions with many colleagues. I am grateful to Max Jensen, Martin Kružík, Marijo Milicevic, Rüdiger Müller, Ricardo H. Nochetto, Christoph Ortner, Alexis Papathanassopoulos, Andreas Prohl, Tomaš Roubíček, Patrick Schreier, Ulisse Stefanelli, and Mirjam Walloth for reading parts of the manuscript and giving me useful hints and comments.

Sören Bartels
Freiburg, September 2013

Part I
Analytical and Numerical Foundations

Chapter 2
Analytical Background

2.1 Variational Model Problems

We describe in this section some model problems that arise in the mathematical description of certain phenomena in continuum mechanics. For justifications of the models, the reader is referred to the textbooks [3, 5, 9, 11].

2.1.1 Deflection of a Membrane

We consider a membrane that occupies the domain $\Omega \subset \mathbb{R}^2$ and is clamped on its boundary $\partial\Omega$. The infinitesimal deflection due to a small vertical force $f \in L^2(\Omega)$ is described by the problem of minimizing the energy functional

$$I(u) = \frac{1}{2} \int\limits_\Omega |\nabla u|^2 \, \mathrm{d}x - \int\limits_\Omega f u \, \mathrm{d}x$$

in the set of functions $u \in H^1(\Omega)$ with $u|_{\partial\Omega} = 0$, cf. Fig. 2.1. The first term in the energy functional I is the *Dirichlet energy*. The Lax–Milgram lemma guarantees the existence of a unique solution that solves the corresponding *Euler–Lagrange equations* which are given by the *Poisson problem* $-\Delta u = f$ in Ω and $u|_{\partial\Omega} = 0$.

2.1.2 Minimal Surfaces

A mathematical model for describing soap films follows from the hypothesis that these minimize surface area. If the surface can be represented as the graph of a function $u : \Omega \to \mathbb{R}$, this leads to the variational problem of minimizing the functional

S. Bartels, *Numerical Methods for Nonlinear Partial Differential Equations*,
Springer Series in Computational Mathematics 47,
DOI 10.1007/978-3-319-13797-1_2

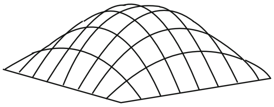


Fig. 2.1 Deflection of a membrane

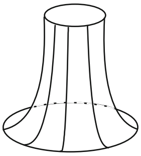


Fig. 2.2 Minimal surface described by a graph

$$I(u) = \int_{\Omega} (1 + |\nabla u|^2)^{1/2} \, \mathrm{d}x$$

in the set of functions $u \in W^{1,p}(\Omega)$ for some appropriate exponent $p \in [1, \infty]$, subject to the boundary condition $u|_{\partial\Omega} = u_{\mathrm{D}}$. If one expects that $|\nabla u| \ll 1$, then the approximation $(1 + |\nabla u(x)|^2)^{1/2} \approx 1 + (1/2)|\nabla u|^2$ justifies using the Dirichlet energy and $p = 2$ to compute solutions. In the general case, the right choice of p is unclear and this is related to the limitations of the model which only applies to situations in which the entire surface is described by a graph. In general this is not true, and large unbounded gradients can occur in minimizing the energy related to functions that do not belong to $W^{1,p}(\Omega)$, cf. Fig. 2.2.

2.1.3 Hyperelastic Materials

Many solid materials behave in an elastic way for a large range of forces, i.e., when a force acts on the body it deforms and when the force stops acting, the body returns to its reference configuration, e.g., the material behaves like a sponge or a network of elastic springs, cf. Fig. 2.3. One can then justify that the actual *deformation* $y : \Omega \to \mathbb{R}^3$ of the body $\Omega \subset \mathbb{R}^3$, due to a force $f : \Omega \to \mathbb{R}^3$, such as gravity minimizes an energy functional of the form

$$I(y) = \int_{\Omega} W(\nabla y) \, \mathrm{d}x - \int_{\Omega} f \cdot y \, \mathrm{d}x$$

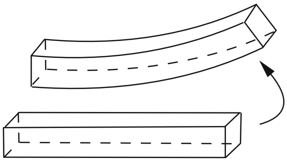


Fig. 2.3 Hyperelastic deformation of a beam

in the set of functions $y \in W^{1,p}(\Omega;\mathbb{R}^3)$ that satisfy boundary conditions $y|_{\Gamma_D} = y_D$ on a subset $\Gamma_D \subset \partial\Omega$. Various physical requirements limit possible choices of W, e.g., since a body cannot be compressed arbitrarily, we require that $W(F) \to \infty$ as $\det F \to 0$ with $\det F > 0$. Moreover, in the absence of a force, and for boundary conditions defined by a rotation of the body, the rotation should also minimize the energy, i.e., $DW(R) = 0$ for every $R \in SO(3)$. These two conditions imply that W cannot be convex. If the body is only slightly displaced from its reference configuration, i.e., $y = \mathrm{id} + u$ with the *displacement* $u : \Omega \to \mathbb{R}^3$ satisfying $|\nabla u| \ll 1$, then with $DW(I) = 0$ we have the approximation

$$W(\nabla[\mathrm{id} + u]) \approx W(I) + \frac{1}{2} D^2 W(I)[\nabla u, \nabla u]$$

which justifies replacing $W(\nabla y)$ by the quadratic expression $(1/2)\mathbb{C}\nabla u : \nabla u$ with the *elasticity tensor* $\mathbb{C} = D^2 W(I)$ and with $A : B$ denoting the inner product of two matrices. The resulting model of *linear elasticity* is important in many applications where only small strains occur.

2.1.4 Obstacle Problems

When the deflection of an elastic membrane is restricted by an *obstacle*, then a constraint has to be included in the above minimization problem, i.e., we seek $u \in H^1(\Omega)$ with $u|_{\partial\Omega} = u_D$, which is a solution of the constrained minimization problem defined by

$$I(u) = \frac{1}{2}\int_\Omega |\nabla u|^2 \, \mathrm{d}x - \int_\Omega f u \, \mathrm{d}x \quad \text{subject to } u \geq \chi \text{ in } \Omega.$$

It is not known in advance and depends on the force and the boundary conditions where the membrane will touch the obstacle described by the function χ. Therefore, determining a *free boundary* that separates the *contact zone* $\mathscr{C} = \{x \in \Omega : u(x) = \chi(x)\}$ from the *noncontact zone* is part of the problem. A model situation is illustrated in Fig. 2.4.

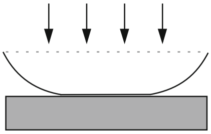


Fig. 2.4 Cross-section of the constrained deflection of a membrane

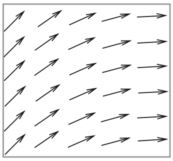


Fig. 2.5 Unit length vector field that may describe the orientation of liquid crystal molecules

2.1.5 Harmonic Maps

A different type of constraint arises in modeling certain liquid crystals and ferromagnets. If the vector field $u : \Omega \to \mathbb{R}^3$ describes the orientation of the rod-like molecules of a liquid crystal or the magnetization of a ferromagnet, then it is natural to impose the pointwise constraint $|u(x)| = 1$ in Ω. In a greatly simplified setting, we then consider the energy functional

$$I(u) = \frac{1}{2} \int_{\Omega} |\nabla u|^2 \,\mathrm{d}x \quad \text{subject to } |u(x)| = 1 \text{ in } \Omega,$$

together with boundary conditions such as $u|_{\Gamma_\mathrm{D}} = u_\mathrm{D}$. Solutions of this minimization problem are called *harmonic maps into the sphere*. Such vector fields can have strong singularities; a smooth unit-length vector field is depicted in Fig. 2.5.

2.1.6 Phase-Field Models

Pointwise constraints in minimization problems are often included by adding a *penalty term* to the energy functional, e.g., by considering

$$I(u) = \frac{1}{2} \int_{\Omega} |\nabla u|^2 \,\mathrm{d}x + \frac{1}{4\varepsilon^2} \int_{\Omega} (|u|^2 - 1)^2 \,\mathrm{d}x$$

with a small parameter $0 < \varepsilon \ll 1$. Deviations of the function $u : \Omega \to \mathbb{R}$ or the vector field $u : \Omega \to \mathbb{R}^3$ from unit length are thus strongly penalized, but in general minimizers will not satisfy $|u| = 1$ everywhere. In the case of a scalar function $u : \Omega \to \mathbb{R}$, this leads to the formation of an *interface* $\Gamma \subset \Omega$ that separates regions in which $u \approx 1$ and $u \approx -1$, cf. Fig. 2.6. The energy functional I with fixed ε occurs in modeling certain *phase transition models* and then the function u describes two different phases, such as a solid and a liquid phase by the values ± 1 and the interface via $\Gamma = \{x \in \Omega : u(x) = 0\}$.

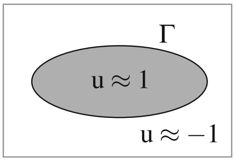


Fig. 2.6 A function that represents the phases in a binary phase separation process

2.1.7 Plate Bending

When the thickness of a hyperelastic body Ω is small, e.g., if $\Omega = \omega \times (-t/2, t/2)$ with a plate thickness $0 < t \ll 1$, then the behavior of the body changes drastically. In such situations it is desirable to treat the body as a two-dimensional object, i.e., to consider an appropriate limit $t \to 0$, described by the deformation of the mid-surface $\omega \subset \mathbb{R}^2$. For a large class of forces f and boundary conditions y_{D}, the minimal value of the three-dimensional energy minimization problem is proportional to t^3 called the bending regime, from which one then can rigorously derive a dimensionally reduced model that describes the deformation $y : \omega \to \mathbb{R}^3$ of the mid-surface ω via the constrained energy functional

$$I^{2D}(y) = \frac{1}{2}\int_{\omega} |D^2 y|^2 \,\mathrm{d}x - \int_{\omega} \tilde{f} \cdot y \,\mathrm{d}x \quad \text{subject to } \nabla y^\top \nabla y = I \text{ in } \Omega$$

together with the boundary conditions $y|_{\Gamma_{\mathrm{D}}} = y_{\mathrm{D}}$ and $\nabla y|_{\Gamma_{\mathrm{D}}} = B_{\mathrm{D}}$. The pointwise constraint on the deformation gradient implies that y is an *isometry*, i.e., that locally, length and angle relations are preserved. The model describes the deformation of a sheet of paper or a piece of cloth, cf. Fig. 2.7.

2.1.8 Crystalline Phase Transitions

For certain crystalline solids the structure of the crystal lattice is temperature-dependent and this enables many important technical applications based on the *shape-memory effect*. The underlying mechanism is that the crystal lattice is highly symmetric, e.g., cubic, for temperatures above a *transition temperature* θ_0 and less structured for low temperatures, e.g., tetragonal, cf. Fig. 2.8. In the low-temperature phase the material can be deformed easily, and in the high-temperature phase it is stiff

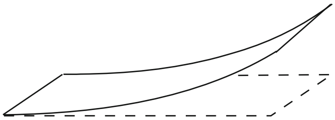


Fig. 2.7 Isometric deformation of a thin elastic sheet

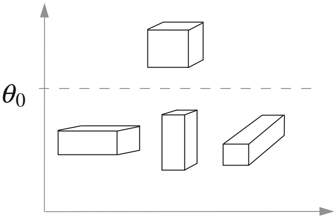


Fig. 2.8 Crystalline phase transition

and tends to return to its reference configuration. If the less structured crystal lattice is described by matrices $F_1, \ldots, F_J \in \mathbb{R}^{3\times 3}$, then a simplified model for the elastic deformation of the material below the transition temperature leads to minimizing the energy functional

$$I(y) = \int_\Omega W(\nabla y)\,\mathrm{d}x \quad \text{with} \quad W(F) = \min_{j=1,\ldots,J} \frac{1}{2}|F - F_j|^2.$$

The *nonconvexity* of the energy density W results in developing oscillations of the deformation gradient ∇y between the values $F_1, \ldots, F_J$. Due to the lack of a characteristic length scale, the oscillations become arbitrarily rapid and a minimizer of I may not exist. This fact requires that the model be appropriately modified in order to capture relevant information about the macroscopic material behavior.

2.1.9 Free-Discontinuity Problems

Many modern applications, including image processing and damage or fracture of materials, require describing certain quantities by discontinuous functions. One approach to their mathematical modeling is to consider a generalized, measure-valued gradient, and this allows us to treat functions that are piecewise smooth with an appropriate discontinuity set, e.g., the characteristic function of a square or a disk. In order to regularize a given noisy image described by its gray values by a function $g : \Omega \to \mathbb{R}$ while preserving its edges, the *total-variation regularized* model seeks a function $u : \Omega \to \mathbb{R}$ that minimizes the functional

$$I(u) = \int_\Omega |Du| + \frac{\alpha}{2}\|u - g\|^2_{L^2(\Omega)}.$$

The second term in the energy functional makes sure that u is close to g, while the first term prohibits certain oscillations, cf. Fig. 2.9. For weakly differentiable functions $u \in W^{1,1}(\Omega)$, the first term coincides with $\|\nabla u\|_{L^1(\Omega)}$, but it is also finite for a large class of discontinuous functions, e.g., if $u = \chi_A$ is the characteristic function of a set $A \subset \Omega$, then $\int_\Omega |Du|$ is the length of ∂A.

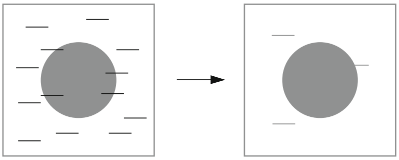


Fig. 2.9 Denoising of a perturbed image

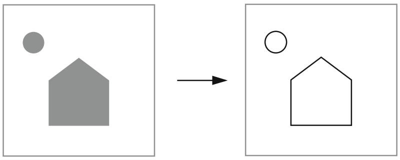


Fig. 2.10 Segmentation of an image

2.1.10 Segmentation Models

Modeling a crack in a material typically requires an explicit description of the crack. Similarly, the problem of detecting certain shapes in images can be described by considering the unknown contour of objects as a separate variable in a model. A simple model problem is defined by the *Mumford–Shah functional*

$$I(u, K) = \frac{1}{2} \int_{\Omega \setminus K} |\nabla u|^2 \,\mathrm{d}x + \mathscr{H}^{d-1}(K) + \frac{\alpha}{2} \|u - g\|^2_{L^2(\Omega)}.$$

Here, $K \subset \overline{\Omega}$ is a closed subset and its $(d-1)$-dimensional Hausdorff measure $\mathscr{H}^{d-1}(K)$ is finite if K belongs to a class of certain lower-dimensional objects such as unions of curves for $d = 2$. A minimizing function $u \in L^2(\Omega)$ has to be weakly differentiable in $\Omega \setminus K$ and close to g, but it may jump and have discontinuities across K. The minimization problem detects contours in a given image g and identifies objects as depicted in Fig. 2.10.

2.1.11 Elastoplasticity

The restoring force or *stress* σ of an elastic spring or rubber band of initial length ℓ that is elongated by an external loading with *strain* $\varepsilon(u) = u'$ is according to *Hooke's law* given by $\sigma = \mathbb{C}\varepsilon(u)$ for a certain range of strains. When σ reaches a critical value σ_y called *yield stress*, then the material behavior changes and a remaining, *plastic* deformation occurs, i.e., after the experiment we observe that the length $\widetilde{\ell}$ of the rubber band is bigger than its initial length ℓ. This is accompanied by a change of the microstructural properties of the material, e.g., of the crystal lattice. Mathematically, this can be described by the requirement that $\sigma \in S = \overline{B_{\sigma_y}(0)}$ and that plastic material behavior occurs if $\sigma \in \partial S$. In this case an increasing strain

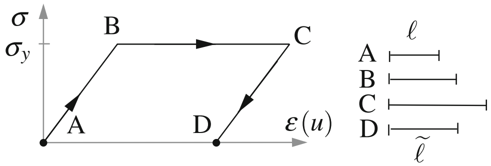


Fig. 2.11 Stress-strain relation for perfectly plastic material behavior; from A to B the material behaves elastically, from B to C a plastic strain occurs leading to a remaining deformation, and from C to D the elastic part of the deformation relaxes

$\varepsilon(u)$ is compensated by a developing *plastic strain* p, i.e., $\sigma = \mathbb{C}(\varepsilon(u) - p)$. The nonlinear stress-strain relation is depicted in Fig. 2.11. With the equilibrium of forces in a quasistatic situation, the process is modeled by the equations

$$-\operatorname{div}\sigma = f, \quad \varepsilon(u) = \mathbb{C}^{-1}\sigma + p, \quad \dot{p} \in \partial I_S(\sigma)$$

with ∂I_S denoting the subdifferential of the indicator functional of S which is the normal cone mapping related to the convex set S, i.e.,

$$\partial I_S(\sigma) = \begin{cases} \{0\} & \text{if } |\sigma| < \sigma_y, \\ \{\alpha\sigma : \alpha \geq 0\} & \text{if } |\sigma| = \sigma_y, \\ \emptyset & \text{if } |\sigma| > \sigma_y. \end{cases}$$

This is a time-dependent formulation and after discretization in time, one is led to solve for every time-step t_k the minimization problem

$$I^k(u,p) = \frac{\sigma_y}{\tau}\int_\Omega |p - p^{k-1}|\,\mathrm{d}x + \frac{1}{2}\int_\Omega |\mathbb{C}^{1/2}(\varepsilon(u) - p)|^2\,\mathrm{d}x - \int_\Omega f \cdot u\,\mathrm{d}x,$$

where p^{k-1} is the solution of the previous time step and subject to time-dependent boundary conditions $u|_{\Gamma_\mathrm{D}} = u_\mathrm{D}(t_k)$. In a more realistic description, deformations that are pure compressions or tensions do not lead to plastic deformations and only $|\operatorname{dev}(\sigma)| \leq \sigma_y$ is required with the *deviator* $\operatorname{dev} A = A - (1/d)(\operatorname{tr} A)I$ of a matrix $A \in \mathbb{R}^{d\times d}$. Moreover, additional variables are included to describe the internal properties of the material.

2.2 Existence of Minimizers

We discuss in this section sufficient and necessary conditions for the existence of minimizers for energy functionals of the form

$$I(u) = \int_{\Omega} W(x, u(x), \nabla u(x))\,\mathrm{d}x$$

in a set of weakly differentiable admissible functions $\mathscr{A} \subset W^{1,p}(\Omega;\mathbb{R}^m)$ for a bounded Lipschitz-domain $\Omega \subset \mathbb{R}^d$. The main idea is to develop an appropriate generalization of the Bolzano–Weierstraß theorem to infinite-dimensional situations. In particular, to deduce compactness properties of bounded sets, it is necessary to work with weak topologies and the usual continuity assumption is replaced by (weak) lower semicontinuity. For further details and more general statements the reader is referred to the textbooks [1, 4, 6, 10].

2.2.1 The Direct Method in the Calculus of Variations

We consider a functional $F : X \to \mathbb{R} \cup \{+\infty\}$ defined on a real, reflexive Banach space X and discuss the existence of minimizers of F.

Definition 2.1 The function $F : X \to \mathbb{R} \cup \{+\infty\}$ is called *weakly lower semicontinuous* if for every sequence $(v_n)_{n\in\mathbb{N}} \subset X$ and $v \in X$ with $v_n \rightharpoonup v$ as $n \to \infty$, i.e., $\phi(v_n) \to \phi(v)$ for every $\phi \in X'$, we have

$$F(v) \le \liminf_{n\to\infty} F(v_n).$$

The validity and the failure of the requirement are illustrated in Fig. 2.12.

Remarks 2.1 (i) The sequence $(F(v_n))_{n\in\mathbb{N}} \subset \mathbb{R} \cup \{+\infty\}$ may be divergent and the definition requires that $F(v)$ be a lower bound for all accumulation points of the sequence.

(ii) In infinite-dimensional spaces, weak lower semicontinuity is a stronger requirement than (strong) lower semicontinuity, i.e., $F(v) \le \liminf_{n\to\infty} F(v_n)$ whenever v_n converges (strongly) to v. Mazur's lemma implies that every convex, (strongly) lower semicontinuous functional is weakly lower semicontinuous. In finite-dimensional spaces the notions of weak and strong lower semicontinuity coincide.

To invoke the fact that according to the Eberlein–Šmuljan theorem every bounded sequence in a reflexive Banach space has a weakly convergent subsequence, we need to assume that the functional F grows outside of bounded sets.

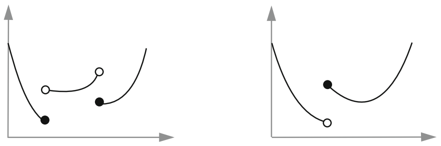


Fig. 2.12 A function that is lower semicontinuous (*left*) and a function that is not lower semicontinuous (*right*)

Definition 2.2 The functional $F : X \to \mathbb{R} \cup \{+\infty\}$ is called *(weakly) coercive* if for every sequence $(v_n)_{n\in\mathbb{N}} \subset X$ with $\|v_n\| \to \infty$, we have $F(v_n) \to \infty$ as $n \to \infty$.

The following theorem can be generalized in several directions. It is formulated in a way that is applicable to many of the variational problems discussed above.

Theorem 2.1 (Direct method in the calculus of variations) *Assume that $F : X \to \mathbb{R} \cup \{+\infty\}$ is weakly lower semicontinuous, coercive, bounded from below, and there exists $v_0 \in X$ with $F(v_0) \in \mathbb{R}$. Then F has a minimizer.*

Proof The proof follows in three steps.
Step 1: Since F is bounded from below, there exists an infimizing sequence $(v_n)_{n\in\mathbb{N}} \subset X$ with $\lim_{n\to\infty} F(v_n) = \inf_{v'\in X} F(v')$.
Step 2: The assumed coercivity of F implies that the sequence $(v_n)_{n\in\mathbb{N}}$ is bounded and therefore, using that X is reflexive, we may extract a weakly convergent subsequence $(v_{n_k})_{k\in\mathbb{N}}$ with weak limit $v \in X$.
Step 3: Due to the weak lower semicontinuity, we have $F(v) \leq \liminf_{k\to\infty} F(v_{n_k})$ and therefore it follows that

$$F(v) \leq \liminf_{k\to\infty} F(v_{n_k}) = \lim_{k\to\infty} F(v_{n_k}), = \inf_{v'\in X} F(v'),$$

i.e., since $F(v) \geq \inf_{v'\in X} F(v')$ we have $F(v) = \inf_{v'\in X} F(v')$ which proves the theorem. □

Remark 2.2 If a variational problem is formulated on a subset $A \subset X$, then we need to impose that A be weakly closed to ensure that the weak accumulation points of a bounded sequence belong to A. This is equivalent to the condition that the *indicator functional* $I_A : X \to \mathbb{R} \cup \{+\infty\}$, defined by $I_A(v) = 0$ for $v \in A$ and $I_A(v) = +\infty$ otherwise, be weakly lower semicontinuous. By Mazur's lemma, it suffices that A be convex and closed.

Examples 2.1 (i) The Dirichlet energy $I(u) = (1/2) \int_\Omega |\nabla u|^2 \,\mathrm{d}x$ is weakly lower semicontinuous since according to a binomial formula, we have

$$\int\limits_\Omega |\nabla u|^2 \,\mathrm{d}x - \int\limits_\Omega |\nabla u_n|^2 + \int\limits_\Omega |\nabla(u-u_n)|^2 \,\mathrm{d}x = 2\int\limits_\Omega \nabla u \cdot \nabla(u-u_n)\,\mathrm{d}x,$$

and if $u_n \rightharpoonup u$ in $H^1(\Omega)$, i.e., $\int_\Omega \nabla u \cdot \nabla(u-u_n)\,\mathrm{d}x \to 0$, this implies that $I(u) \leq \liminf_{n\to\infty} I(u_n)$. The coercivity of I follows from a Poincaré inequality.
(ii) Simple examples such as Weierstraß' example show that not every minimization problem has a solution. By constructing an infimizing sequence consisting of Lipschitz continuous functions, one can verify that the functional

$$I(y) = \int\limits_{(-1,1)} (xy'(x))^2 \,\mathrm{d}x$$

has no continuous minimizer subject to the boundary conditions $y(-1) = -1$ and $y(1) = 1$ although I is weakly lower semicontinuous and bounded from below. Similarly, one can show that there is no differentiable minimizer of

$$I(u) = \int_{(0,1)} (|u'(x)|^2 - 1)^2 \, \mathrm{d}x + \int_{\Omega} |u(x)|^4 \, \mathrm{d}x,$$

but I is coercive and bounded from below.

2.2.2 Sobolev Spaces

To investigate functionals that are defined as integrals of integrands applied to functions and their derivatives, we always consider a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ and recall that a function $u \in L^1(\Omega)$ is called *weakly differentiable* if there exists a vector field $G \in L^1(\Omega; \mathbb{R}^d)$ such that

$$\int_{\Omega} u \,\mathrm{div}\, \varphi \, \mathrm{d}x = - \int_{\Omega} G \cdot \varphi \, \mathrm{d}x$$

for all smooth, compactly supported vector fields $\varphi \in C_0^\infty(\Omega; \mathbb{R}^d)$. We call G the *weak gradient* of u and write $\nabla u = G$. We then define

$$\begin{aligned} W^{1,p}(\Omega; \mathbb{R}^m) = \{u = (u_1, \dots, u_m) \in L^p(\Omega; \mathbb{R}^m) : \\ \nabla u_j \in L^p(\Omega; \mathbb{R}^d), \ j = 1, 2, \dots, m\}, \end{aligned}$$

which is a Banach space for the norm $\|u\|_{W^{1,p}} = (\|u\|_{L^p}^p + \|\nabla u\|_{L^p}^p)^{1/p}$. For a closed subset $\Gamma_{\mathrm{D}} \subset \partial\Omega$ with positive surface measure, we set

$$W_{\mathrm{D}}^{1,p}(\Omega; \mathbb{R}^m) = \{v \in W^{1,p}(\Omega; \mathbb{R}^m) : v|_{\Gamma_{\mathrm{D}}} = 0\}$$

and write $W_0^{1,p}(\Omega; \mathbb{R}^m)$ if $\Gamma_{\mathrm{D}} = \partial\Omega$. We recall some important facts about Sobolev spaces. For $p = 2$, we have that $W^{1,2}(\Omega; \mathbb{R}^m)$ is a Hilbert space denoted by $H^1(\Omega; \mathbb{R}^m)$, and analogously $H_{\mathrm{D}}^1(\Omega; \mathbb{R}^m)$ stands for $W_{\mathrm{D}}^{1,2}(\Omega; \mathbb{R}^m)$.

Remarks 2.3 (i) If $p < d$, then the *embedding* $W^{1,p}(\Omega; \mathbb{R}^m) \to L^q(\Omega; \mathbb{R}^m)$ is continuous for $1 \le q \le p^*$ with the Sobolev conjugate exponent $p^* = pd/(d-p)$, i.e., $\|u\|_{L^q(\Omega)} \le c\|u\|_{W^{1,p}(\Omega)}$ for a constant $c > 0$ and every $u \in W^{1,p}(\Omega; \mathbb{R}^m)$. If $p = d$, then this is true for every $1 \le q < \infty$. If $p > d$, then the embedding $W^{1,p}(\Omega; \mathbb{R}^m) \to C(\overline{\Omega}; \mathbb{R}^m)$ is continuous.

(ii) The embeddings are *compact*, i.e., whenever $u_j \rightharpoonup u$ in $W^{1,p}(\Omega;\mathbb{R}^m)$, then it follows that $u_j \to u$ in $L^q(\Omega;\mathbb{R}^m)$, provided that $1 \le q < p^*$ if $p > d$ and $1 \le q < \infty$ if $p = d$ and $1 \le q \le \infty$ if $p > d$.
(iii) The subset $C^\infty(\overline{\Omega};\mathbb{R}^m) \cap W^{1,p}(\Omega;\mathbb{R}^m)$ is *dense* and $W^{1,p}(\Omega;\mathbb{R}^m)$ is *separable* if $1 \le p < \infty$. For $1 < p < \infty$ the space $W^{1,p}(\Omega;\mathbb{R}^m)$ is reflexive.
(iv) For $1 \le p \le \infty$ there exists a bounded linear operator $\operatorname{tr} : W^{1,p}(\Omega;\mathbb{R}^m) \to L^p(\Omega;\mathbb{R}^m)$ called the *trace operator* such that $\operatorname{tr} u = u|_{\partial\Omega}$ for every $u \in C(\overline{\Omega};\mathbb{R}^m)$. We use the notation $u|_{\Gamma_{\rm D}}$ to denote the restriction of the trace to a subset $\Gamma_{\rm D}$ of $\partial\Omega$.
(v) For $1 \le p \le \infty$, $p' = p/(p-1)$, $v \in W^{1,p}(\Omega)$, and $w \in W^{1,p'}(\Omega;\mathbb{R}^d)$, we have *Green's* or the *integration-by-parts formula*

$$\int_\Omega v \operatorname{div} w \,\mathrm{d}x + \int_\Omega \nabla v \cdot w \,\mathrm{d}x = \int_{\partial\Omega} \operatorname{tr}(v)\operatorname{tr}(w)\cdot n \,\mathrm{d}s,$$

where n denotes the outer unit normal to Ω on $\partial\Omega$.
(vi) *Poincaré inequalities* bound the norm $\|u\|_{W^{1,p}}$ for $1 \le p \le \infty$ by the semi-norm $|u|_{W^{1,p}} = \|\nabla u\|_{L^p}$ for $u \in W^{1,p}(\Omega;\mathbb{R}^m)$, i.e., $\|u\|_{W^{1,p}} \le C_P |u|_{W^{1,p}}$, provided that $u|_{\Gamma_{\rm D}} = 0$ for a closed set $\Gamma_{\rm D} \subset \partial\Omega$ with positive surface measure or $\int_\Omega u \,\mathrm{d}x = 0$.
(vii) The closure of the set $C_0^\infty(\Omega;\mathbb{R}^m)$ with respect to the norm $\|\cdot\|_{W^{1,p}}$ coincides with the space $W_0^{1,p}(\Omega;\mathbb{R}^m)$ if $1 \le p < \infty$.
(viii) For $1 \le p \le \infty$, we have $u_n \rightharpoonup u$ in $W^{1,p}(\Omega;\mathbb{R}^m)$ if and only if $u_n \to u$ in $L^p(\Omega;\mathbb{R}^m)$ and $\nabla u_n \rightharpoonup \nabla u$ in $L^p(\Omega;\mathbb{R}^m)$ as $n \to \infty$.

For $k \ge 2$ the higher-order Sobolev spaces $W^{k,p}(\Omega;\mathbb{R}^m)$ are defined inductively, i.e., $W^{k,p}(\Omega;\mathbb{R}^m) = \{v \in W^{1,p}(\Omega;\mathbb{R}^m) : \nabla v \in W^{k-1,p}(\Omega;\mathbb{R}^{m\times d})\}$. A multiindex notation is used to abbreviate higher order partial derivatives, i.e., for $\alpha \in \mathbb{N}_0^d$, we denote

$$\partial^\alpha u = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \dots \partial x_d^{\alpha_d}},$$

where $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_d$. The k-th derivative of u is the vector containing all weak partial derivatives of order k, i.e., $D^k u = (\partial^\alpha u)_{|\alpha| \le k}$. As above, we write $H^k(\Omega;\mathbb{R}^m)$ if $p = 2$.

2.2.3 Integral Functionals

For integral functionals defined on Sobolev spaces of scalar functions in $W^{1,p}(\Omega)$, precise conditions on an integrand that imply weak lower semicontinuity are known. For vector fields in spaces $W^{1,p}(\Omega;\mathbb{R}^m)$, the conditions of the following theorem are only sufficient.

Theorem 2.2 (Weak lower semicontinuity of integral functionals) *Let* $1 \le p < \infty$ *and assume that* $W : \mathbb{R}^{m\times d} \to \mathbb{R}$ *is continuous with* $|W(A)| \le c(1 + |A|^p)$. *If* W *is* *convex, then the functional*

$$I(u) = \int_{\Omega} W(\nabla u)\,\mathrm{d}x$$

is weakly lower semicontinuous on $W^{1,p}(\Omega;\mathbb{R}^m)$*. Conversely, if* $m = 1$ *and* I *is weakly lower semicontinuous on* $W^{1,p}(\Omega)$*, then* W *is convex.*

Proof (1) For a simpler proof we assume that W is convex and continuously differentiable with $|DW(A)| \le c'(1+|A|^{p-1})$. We then have $W(B) \ge W(A) + DW(A)\cdot(B-A)$ for all $A, B \in \mathbb{R}^{m\times d}$. Due to the estimate $(a+b)^s \le 2^{s-1}(a^s+b^s)$ for all $a, b \in \mathbb{R}$ and $s \ge 1$, we thus have for every $u \in W^{1,p}(\Omega;\mathbb{R}^m)$ that

$$\int_{\Omega} |DW(\nabla u)|^{p/(p-1)}\,\mathrm{d}x \le c\int_{\Omega} (1+|\nabla u|^{p-1})^{p/(p-1)}\,\mathrm{d}x \le c\int_{\Omega}(1+|\nabla u|)^p\,\mathrm{d}x,$$

i.e., $DW(\nabla u) \in L^{p'}(\Omega;\mathbb{R}^{m\times d})$. If $u_n \rightharpoonup u$ in $W^{1,p}(\Omega;\mathbb{R}^m)$ as $n\to\infty$, then we have $\nabla(u_n-u)\rightharpoonup 0$ in $L^p(\Omega;\mathbb{R}^{m\times d})$. Using this in the estimate

$$\int_{\Omega} W(\nabla u_n)\,\mathrm{d}x \ge \int_{\Omega} W(\nabla u)\,\mathrm{d}x + \int_{\Omega} DW(\nabla u)\cdot\nabla(u_n-u)\,\mathrm{d}x,$$

we observe that the second term on the right-hand side converges to 0 as $n\to\infty$. This implies that $\liminf_{n\to\infty} I(u_n) \ge I(u)$, i.e., that I is weakly lower semicontinuous. A more general proof which avoids the assumption that W is continuously differentiable employs Mazur's lemma.

(2) To prove the converse implication, we let $A, B \in \mathbb{R}^d$, $\theta \in [0,1]$, set $F = \theta A + (1-\theta)B$, and define $u_F(x) = Fx$ for $x \in \Omega$. We assume that for every $\varepsilon > 0$ there exists a function $v_\varepsilon \in W^{1,\infty}(\Omega)$, such that $\|\nabla v_\varepsilon\|_{L^\infty(\Omega)} \le c$ independently of ε, $\|v_\varepsilon - u_F\|_{L^\infty(\Omega)} \le c\varepsilon$, and for $\Omega_X^\varepsilon = \{x\in\Omega : \nabla v_\varepsilon(x) = X\}$ with $X \in \{A, B\}$ we have

$$\mathscr{L}^d(\Omega_A^\varepsilon) \le \theta\mathscr{L}^d(\Omega) + c\varepsilon,\ \mathscr{L}^d(\Omega_B^\varepsilon) \le (1-\theta)\mathscr{L}^d(\Omega) + c\varepsilon,$$
$$\mathscr{L}^d(\Omega\setminus(\Omega_A^\varepsilon\cup\Omega_B^\varepsilon)) \le c\varepsilon.$$

The construction of such a function will be discussed in the subsequent lemma. For every $n\in\mathbb{N}$, set $\varepsilon_n = 1/n$ and let $u_n = v_{\varepsilon_n}$. We then have

$$\begin{aligned}\int_{\Omega} W(\nabla u_n)\,\mathrm{d}x &= \mathscr{L}^d(\Omega_A^{1/n})W(A) + \mathscr{L}^d(\Omega_B^{1/n})W(B) + \int_{\Omega\setminus(\Omega_A^{1/n}\cup\Omega_B^{1/n})} W(\nabla u_n)\,\mathrm{d}x \\ &\le \theta\mathscr{L}^d(\Omega)W(A) + (1-\theta)\mathscr{L}^d(\Omega)W(B) + c/n.\end{aligned}$$

Since $u_n \rightharpoonup u_F$ in $W^{1,p}(\Omega)$ as $n \to \infty$ and $\nabla u_F = F$ in Ω we deduce with the assumed weak lower semicontinuity of I that

$$\begin{aligned}\mathscr{L}^d(\Omega)W(F) = \int_\Omega W(\nabla u_F)\,\mathrm{d}x &\le \liminf_{n\to\infty} \int_\Omega W(\nabla u_n)\,\mathrm{d}x \\ &\le \theta\mathscr{L}^d(\Omega)W(A) + (1-\theta)\mathscr{L}^d(\Omega)W(B),\end{aligned}$$

i.e., that W is convex. □

Remarks 2.4 (i) For integrands of the form $W(x, u(x), \nabla u(x))$ one needs to assume that W is a *Carathéodory function*, i.e., that for all $(s, A) \in \mathbb{R}^m \times \mathbb{R}^{m\times d}$, the mapping $x \mapsto W(x, s, A)$ is measurable and for almost every $x \in \Omega$, the mapping $(s, A) \mapsto W(x, s, A)$ is continuous. A sufficiency requirement for weak lower semicontinuity is then in addition to certain growth conditions that $A \mapsto W(x, s, A)$ be convex for almost every $x \in \Omega$ and all $s \in \mathbb{R}^m$.
(ii) The functional $J(u) = \int_\Omega j(x, u(x))\,\mathrm{d}x$ is weakly continuous on $W^{1,p}(\Omega; \mathbb{R}^m)$ if, e.g., $|j(x, s)| \le a(x) + |s|^q$ with $a \in L^1(\Omega)$ and $1 \le q \le p^*$, i.e., we have $J(u_n) \to J(u)$ whenever $u_n \rightharpoonup u$ as $n \to \infty$. Moreover, in this case we have that if $I : W^{1,p}(\Omega; \mathbb{R}^m) \to \mathbb{R}$ is weakly lower semicontinuous then $I + J$ is also weakly lower semicontinuous.
(iii) In the vectorial case $m > 1$, convexity is not a necessary condition. Sufficient for weak lower semicontinuity is the weaker notion of *polyconvexity* which requires that there exist a convex function $\widehat{W}$ with $W(A) = \widehat{W}(T(A))$, where $T(A)$ is the vector that contains the determinants of all square submatrices of A, e.g., $T(A) = (A, \det(A))$ if $d = m = 2$. Necessary and sufficient for weak lower semicontinuity on $W^{1,p}(\Omega; \mathbb{R}^m)$ is *quasiconvexity* which requires that

$$W(A) \le \inf_{\varphi\in W_0^{1,p}(\Omega;\mathbb{R}^m)} \frac{1}{\mathscr{L}^d(\Omega)} \int_\Omega W(A + \nabla\varphi)\,\mathrm{d}x,$$

i.e., that the affine function $u(x) = Ax + b$ be minimal for I in the set of all functions in $W^{1,p}(\Omega; \mathbb{R}^m)$ satisfying the same affine boundary conditions. For $m = 1$, quasiconvexity is equivalent to convexity.

We next show in a more general setting how the function v_ε used in the proof of Theorem 2.2 can be constructed.

Lemma 2.1 (Compatible gradients) *Let $A, B \in \mathbb{R}^{m\times d}$ with* $\operatorname{rank}(B - A) = 1$, *i.e., $A - B = \beta \otimes \alpha$ for $\beta \in \mathbb{R}^m$ and $\alpha \in \mathbb{R}^d$ with $|\alpha| = 1$. Let $\theta \in [0, 1]$, set $F = \theta A + (1-\theta)B$, and define $u_F(x) = Fx$ for $x \in \Omega$. For every $\varepsilon > 0$, there exists a Lipschitz continuous function $v_\varepsilon \in W^{1,\infty}(\Omega; \mathbb{R}^m)$ such that $v_\varepsilon = u_F$ on $\partial\Omega$, $\|\nabla v_\varepsilon\|_{L^\infty(\Omega)} \le c$, $\|v_\varepsilon - u_F\|_{L^\infty(\Omega)} \le c\varepsilon$, and for $\Omega_X^\varepsilon = \{x \in \Omega : \nabla v_\varepsilon(x) = X\}$ with $X \in \{A, B\}$, we have*

$$\mathscr{L}^d(\Omega_A^\varepsilon) \le \theta \mathscr{L}^d(\Omega) + c\varepsilon, \ \mathscr{L}^d(\Omega_B^\varepsilon) \le (1-\theta)\mathscr{L}^d(\Omega) + c\varepsilon,$$
$$\mathscr{L}^d(\Omega \setminus (\Omega_A^\varepsilon \cup \Omega_B^\varepsilon)) \le c\varepsilon.$$

Proof By replacing A and B by $A - F$ and $B - F$, we may assume that $F = 0$. We define $\Omega_\varepsilon = \{x \in \Omega : \operatorname{dist}(x, \partial\Omega) > \varepsilon\}$ and choose a cut-off function $\eta_\varepsilon \in W^{1,\infty}(\Omega)$ such that $\eta_\varepsilon|_{\Omega_\varepsilon} = 1$, $|\nabla \eta_\varepsilon(x)| \le c/\varepsilon$ for almost every $x \in \Omega$, and $\eta_\varepsilon|_{\partial\Omega} = 0$. For $x \in \mathbb{R}^d$ we define

$$\widetilde{v}_\varepsilon(x) = Ax - \beta \int_0^{x\cdot\alpha} \widetilde{\chi}_{(\theta,1)}(t/\varepsilon)\,\mathrm{d}t$$

with the 1-periodic function $\widetilde{\chi} : \mathbb{R} \to \mathbb{R}$ given by $\widetilde{\chi}|_{(0,\theta)} = 0$ and $\widetilde{\chi}|_{(\theta,1)} = 1$. The function $\widetilde{v}_\varepsilon$, for $k\varepsilon \le x \cdot \alpha \le (k+1)\varepsilon$ with $k \in \mathbb{Z}$, satisfies that

$$\nabla \widetilde{v}_\varepsilon(x) = \begin{cases} A & \text{if } k\varepsilon < x \cdot \alpha \le (k+\theta)\varepsilon, \\ B & \text{if } (k+\theta)\varepsilon < x \cdot \alpha \le (k+1)\varepsilon. \end{cases}$$

We finally set

$$v_\varepsilon = \eta_\varepsilon \widetilde{v}_\varepsilon$$

and, noting that $\|\widetilde{v}_\varepsilon\|_{L^\infty(\Omega)} \le c\varepsilon$, we verify that this function satisfies the requirements of the lemma. The construction is depicted in Fig. 2.13. □

Remark 2.5 One can show that a nontrivial function $u \in W^{1,\infty}(\Omega; \mathbb{R}^m)$ with $\nabla u(x) \in \{A, B\}$ almost everywhere in Ω can only exist if $\operatorname{rank}(B - A) = 1$. The necessity of this condition follows from the continuity of the tangential gradient along the interface Γ that separates regions of constant gradients.

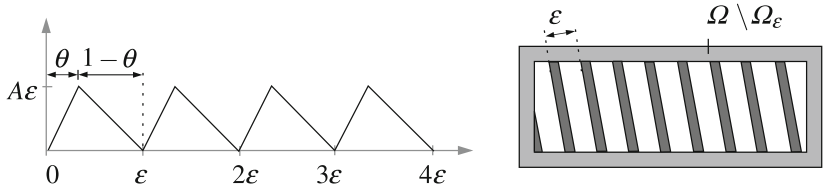


Fig. 2.13 The oscillating function $\widetilde{v}_\varepsilon$ for $d = m = 1$ with an average gradient 0 constructed in the proof of Lemma 2.1 (*left*); the gradient of the function $v_\varepsilon = \eta_\varepsilon \widetilde{v}_\varepsilon$ oscillates between the values A and B away from the boundary (*right*)

2.2.4 Existence and Properties of Minimizers

Certain generalizations of the weak lower semicontinuity results discussed above imply the following existence result for minimization problems on Sobolev spaces.

Theorem 2.3 (Existence) *Let $1 < p < \infty$ and let $W : \Omega \times \mathbb{R}^m \times \mathbb{R}^{m\times d} \to \mathbb{R}$ be a Carathéodory function, such that $W(x, s, A) \ge a(x) + b|s|^q + c|A|^p$ for $1 \le q < p$ and $a \in L^1(\Omega)$, $b, c > 0$ for almost every $x \in \Omega$, and all $(s, A) \in \mathbb{R}^m \times \mathbb{R}^{m\times d}$, and such that $A \mapsto W(x, s, A)$ is quasiconvex for almost every $x \in \Omega$ and all $s \in \mathbb{R}^m$. Assume that there exists $u_0 \in W^{1,p}(\Omega; \mathbb{R}^m)$ with $W(x, u_0(x), \nabla u_0(x)) \in L^1(\Omega)$. Then there exists a minimizer $u \in W^{1,p}(\Omega)$ for*

$$I(u) = \int_\Omega W(x, u(x), \nabla u(x))\, \mathrm{d}x. \tag{2.1}$$

Proof The result follows from the direct method in the calculus of variations and we refer the reader to [4] for a complete proof. □

Remark 2.6 The existence result can be generalized in many directions. If $b = 0$ is allowed, then boundary conditions have to be imposed on a subset $\Gamma_\mathrm{D} \subset \partial\Omega$ to guarantee coercivity of I.

Theorem 2.4 (Uniqueness) *If the mapping $(s, A) \mapsto W(x, s, A)$ is* strictly convex *for almost every $x \in \Omega$, i.e., if for distinct pairs $(s_1, A_1), (s_2, A_2) \in \mathbb{R}^m \times \mathbb{R}^{m\times d}$, we have*

$$\frac{1}{2}W(x, s_1, A_1) + \frac{1}{2}W(x, s_2, A_2) > W\Big(x, \frac{s_1 + s_2}{2}, \frac{A_1 + A_2}{2}\Big),$$

then there exists at most one minimizer of the functional I defined through W as in (2.1).

Proof Suppose $u_1, u_2 \in W^{1,p}(\Omega; \mathbb{R}^m)$ are distinct minimimzers of I. Then the strict convexity of W implies that

$$\frac{1}{2}I(u_1) + \frac{1}{2}I(u_2) > I\Big(\frac{u_1 + u_2}{2}\Big),$$

which contradicts the assumption that u_1 and u_2 are minimizers. □

As in finite-dimensional situations it is desirable to formulate necessary conditions for minimizers. We restrict to the scalar case $m = 1$ and a simple form of the energy density W.

Theorem 2.5 (Euler–Lagrange equations) *Assume that $W : \Omega \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ is given by $W(x, s, A) = W_0(A) - f(x)s$ with $f \in L^{p'}(\Omega)$ and $W_0 \in C^1(\mathbb{R}^d)$ such that $|DW_0(A)| \le c'(1 + |A|^{p-1})$ for all $A \in \mathbb{R}^d$. If $u \in W^{1,p}(\Omega)$ minimizes I*

among functions in $u_0 + W_0^{1,p}(\Omega)$ *for some* $u_0 \in W^{1,p}(\Omega)$ *and* $1 \le p < \infty$*, then* u *solves the* Euler–Lagrange equations

$$\int_\Omega DW_0(\nabla u) \cdot \nabla v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x$$

for all $v \in W_0^{1,p}(\Omega)$.

Proof Let $v \in W_0^{1,p}(\Omega)$ be fixed and consider the function $\psi(r) = I(u + rv)$, $r \in \mathbb{R}$, which has a minimum at $r = 0$. For $r \ne 0$ we have

$$\begin{aligned}\frac{1}{r}(\psi(r) - \psi(0)) &= \frac{1}{r}(I(u+rv) - I(u)) \\ &= \frac{1}{r}\Big[\int_\Omega W_0(\nabla(u+rv))\,\mathrm{d}x - \int_\Omega f(u+rv)\,\mathrm{d}x \\ &\quad - \int_\Omega W_0(\nabla u)\,\mathrm{d}x + \int_\Omega f u\,\mathrm{d}x\Big] \\ &= \frac{1}{r}\int_\Omega W(\nabla(u+rv)) - W_0(\nabla u)\,\mathrm{d}x - \int_\Omega f v\,\mathrm{d}x.\end{aligned}$$

We define

$$M^r(x) = \frac{1}{r}\big(W_0(\nabla(u+rv)(x)) - W_0(\nabla u(x))\big)$$

and note that $M^r(x) \to DW_0(\nabla u(x)) \cdot \nabla v(x)$ as $r \to 0$ for almost every $x \in \Omega$. To pass to the limit of the integrals with Lebesgue's dominated convergence theorem, we aim at constructing an r-independent, integrable upper bound for $M^r(x)$. We consider $0 < r \le 1$ and note that the fundamental theorem of calculus implies that

$$M^r(x) = \frac{1}{r}\int_0^r \frac{d}{dt} W_0(\nabla(u+tv)(x))\,\mathrm{d}t = \frac{1}{r}\int_0^r DW_0(\nabla(u+rv)(x)) \cdot \nabla v(x)\,\mathrm{d}t.$$

Incorporating the assumed upper bound for DW_0, we have for $t \in (0, r)$ that

$$|M^r(x)| \le c(1 + |\nabla u(x)|^{p-1} + |\nabla v(x)|^{p-1})|\nabla v(x)|$$

and it follows with Hölder's inequality that the right-hand side belongs to $L^1(\Omega)$. We may therefore pass to the limit under the integral and have

$$\frac{1}{r}\big(I(u+rv) - I(u)\big) \to \int_\Omega DW(\nabla u) \cdot \nabla v\,\mathrm{d}x - \int_\Omega f v\,\mathrm{d}x.$$

Since $r = 0$ is minimal for $\psi(r)$, it follows that the right-hand side is nonnegative. Repeating the argument with v replaced by $-v$ leads to the assertion. □

Remark 2.7 The strong form of the Euler–Lagrange equations follows from the *fundamental lemma* in the calculus of variations which asserts that whenever we are given $g \in L^1(\Omega)$ with

$$\int_\Omega g\varphi \,\mathrm{d}x = 0$$

for all $\varphi \in C_0^\infty(\Omega)$, then $g = 0$ almost everywhere in Ω. Therefore, we have

$$-\operatorname{div} DW_0(\nabla u) = f \text{ in } \Omega, \quad u|_{\partial\Omega} = u_0|_{\partial\Omega}.$$

More generally, for appropriate functions W, one can derive the partial differential equation

$$-\operatorname{div} \frac{\partial W}{\partial A}\big(x, u(x), \nabla u(x)\big) + \frac{\partial W}{\partial s}\big(x, u(x), \nabla u(x)\big) = 0.$$

If $W_0 \in C^1(\mathbb{R}^d)$ is convex, then the Euler-Lagrange equations also define a sufficient condition for optimality.

Theorem 2.6 (Sufficiency) *Assume that the assumptions of Theorem 2.5 are satisfied and that W_0 is convex. Suppose that $u \in u_0 + W_0^{1,p}(\Omega)$ satisfies*

$$\int_\Omega DW_0(\nabla u) \cdot \nabla v \,\mathrm{d}x = \int_\Omega f v \,\mathrm{d}x$$

for all $v \in W_0^{1,p}(\Omega)$. Then u is minimal for I for functions in $u_0 + W_0^{1,p}(\Omega)$.

Proof Let $v \in W^{1,p}(\Omega)$. Since W_0 is convex and continuously differentiable, we have

$$W_0(\nabla(u+v)(x)) \geq W_0(\nabla u(x)) + DW_0(\nabla u(x)) \cdot \nabla v(x)$$

for almost every $x \in \Omega$. This and the Euler–Lagrange equations imply that

$$\begin{aligned} I(u+v) &= \int_\Omega W_0(\nabla(u+v)) \,\mathrm{d}x - \int_\Omega f(u+v) \,\mathrm{d}x \\ &\geq \int_\Omega W_0(\nabla u) \,\mathrm{d}x + \int_\Omega DW_0(\nabla u) \cdot \nabla v \,\mathrm{d}x - \int_\Omega f u \,\mathrm{d}x - \int_\Omega f v \,\mathrm{d}x \\ &= \int_\Omega W_0(\nabla u) \,\mathrm{d}x - \int_\Omega f u \,\mathrm{d}x = I(u), \end{aligned}$$

i.e., u is minimal for I. □

Remarks 2.8 (i) Regularity of minimizers, e.g., $u \in H^2(\Omega)$, can be proved if the boundary of Ω is C^2-regular and W is *strongly convex*, i.e., $D^2W(A)[B, B] \geq c|B|^2$ for all $A, B \in \mathbb{R}^{m\times d}$ and $c > 0$.
(ii) If the minimization problem involves a constraint, such as $G(u(x)) = 0$ for almost every $x \in \Omega$ with a continuously differentiable function $G : \mathbb{R}^m \to \mathbb{R}$, then one can formally consider a saddle point problem to derive the Euler–Lagrange equations, e.g., for $W(A) = |A|^2/2$, the problem

$$\inf_{u\in H_0^1(\Omega)} \sup_{\lambda\in L^q(\Omega)} \frac{1}{2}\int_\Omega |\nabla u|^2 \,\mathrm{d}x + \int_\Omega \lambda G(u)\,\mathrm{d}x.$$

The optimality conditions are with $g = DG$ given by

$$\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x + \int_\Omega \lambda g(u)\cdot v\,\mathrm{d}x = 0, \qquad \int_\Omega \mu G(u)\,\mathrm{d}x = 0$$

for all $v \in H_0^1(\Omega; \mathbb{R}^\ell)$ and all $\mu \in L^q(\Omega)$. The unknown variable λ is the *Lagrange multiplier* associated to the constraint $G(u(x)) = 0$ for almost every $x \in \Omega$.
(iii) On the part $\partial\Omega \setminus \Gamma_\mathrm{D}$ where no *Dirichlet boundary conditions* $u|_{\Gamma_\mathrm{D}} = u_\mathrm{D}$ are imposed, the homogeneous *Neumann boundary conditions* $DW_0(\nabla u)\cdot n = 0$ are satisfied. Inhomogeneous Neumann conditions can be specified through a function $g \in L^q(\Gamma_\mathrm{N}; \mathbb{R}^m)$ and a corresponding contribution to the energy functional, e.g.,

$$I(u) = \int_\Omega W_0(\nabla u(x))\,\mathrm{d}x - \int_{\Gamma_\mathrm{N}} gu\,\mathrm{d}s.$$

(iv) The Euler–Lagrange equations define an operator $L : W^{1,p}(\Omega) \to W^{1,p}(\Omega)'$ and we look for $u \in W^{1,p}(\Omega)$ with $L(u) = b$ for a given right-hand side $b \in W^{1,p}(\Omega)'$. Under certain monotonicity conditions on L, the existence of solutions for this equation can be established with the help of discretizations and fixed-point theorems. This is of importance when the partial differential equation is not related to a minimization problem.

2.3 Gradient Flows

The direct method in the calculus of variations provides existence results for global minimizers of functionals but its proof is nonconstructive. In practice, the most robust methods to find stationary points are steepest descent methods. These can often be regarded as discretizations of time-dependent problems. To understand the stability and convergence properties of descent methods, it is important and insightful to analyze the corresponding continuous problems. In finite-dimensional situations we may think of a function $V : \mathbb{R}^n \to \mathbb{R}$ and the ordinary differential equation

$$y' = -\nabla V(y), \quad y(0) = y_0.$$

If $V \in C^2(\mathbb{R}^n)$, then the Picard–Lindelöf theorem guarantees the existence of a unique local solution $y : (-\widehat{T}, \widehat{T}) \to \mathbb{R}^n$. Taking the inner product of the differential equation with y' and using the chain rule to verify $\nabla V(y(t)) \cdot y'(t) = (V \circ y)'$ we have after integration over $(0, T')$

$$\int\limits_0^{T'} |y'|^2 \,\mathrm{d}t + V(y(T')) = V(y(0)).$$

This is called an *energy law* and shows that the function $t \mapsto V(y(t))$ is decreasing. Since the evolution becomes stationary if $\nabla V(y(t)) = 0$, this allows us to find critical points of V with small energy. It is the aim of this section to justify gradient flows for functionals on infinite-dimensional spaces. For more details on this subject, we refer the reader to the textbooks [2, 6–8].

2.3.1 Differentiation in Banach Spaces

We consider a Banach space X and a functional $I : X \to \mathbb{R}$.

Definition 2.3 (a) We say that I is *Gâteaux-differentiable* at $v_0 \in X$ if for all $h \in X$ the limit

$$\delta I(v_0, h) = \lim_{s \to 0} \frac{I(v_0 + sh) - I(v_0)}{s}$$

exists and the mapping $DI(v_0) : X \to \mathbb{R}$, $h \mapsto \delta I(v_0, h)$ is linear and bounded.
(b) We say that I is *Fréchet-differentiable* at $v_0 \in X$ if there exist a bounded linear operator $A : X \to \mathbb{R}$ and a function $\varphi : \mathbb{R} \to \mathbb{R}$ with $\lim_{s \to 0} \varphi(s)/s = 0$ such that

$$I(v_0 + h) - I(v_0) = Ah + \varphi(\|h\|_X).$$

In this case we define $DI(v_0) = A$.

Remark 2.9 If I is Gâteaux-differentiable at every point in a neighborhood of v_0 and DI is continuous at v_0, then I is Fréchet-differentiable at v_0.

The gradient of a functional is the Riesz representative of the Fréchet derivative with respect to a given scalar product.

Definition 2.4 Let H be a Hilbert space such that X is continuously embedded in H. If I is Fréchet-differentiable at $v_0 \in X$ with $DI(v_0) \in H'$, then the *H-gradient* $\nabla_H I(v_0) \in H$ is defined by

$$(\nabla_H I(v_0), v)_H = DI(v_0)[v]$$

for all $v \in H$.

Example 2.2 For $X = H_0^1(\Omega)$ and $I(u) = (1/2)\int_\Omega |\nabla u|^2 \, \mathrm{d}x$, we have

$$\begin{aligned} I(u+sv) - I(u) &= \frac{1}{2}\int_\Omega |\nabla u|^2 + 2s\nabla u \cdot \nabla v + s^2 |\nabla v|^2 \, \mathrm{d}x - \frac{1}{2}\int_\Omega |\nabla u|^2 \, \mathrm{d}x \\ &= s\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x + \frac{s^2}{2}\int_\Omega |\nabla v|^2 \, \mathrm{d}x \end{aligned}$$

and I is Fréchet differentiable with $DI(u)[v] = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x$. For $H = X = H_0^1(\Omega)$ with scalar product $(v, w)_{H_0^1} = \int_\Omega \nabla v \cdot \nabla w \, \mathrm{d}x$, we thus have $\nabla_{H_0^1} I(u) = u$. If $u \in H^2(\Omega) \cap H_0^1(\Omega)$, then Green's formula shows that

$$DI(u)[v] = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x = \int_\Omega (-\Delta u) v \, \mathrm{d}x,$$

so that $DI(u)$ is a bounded linear functional on $L^2(\Omega)$. For $H = L^2(\Omega)$ with scalar product $(v, w) = \int_\Omega vw \, \mathrm{d}x$, we therefore have $\nabla_{L^2} I(u) = -\Delta u$.

Remark 2.10 The Euler–Lagrange equations for $I(u) = \int_\Omega W(x, u, \nabla u) \, \mathrm{d}x$ in the strong form corresponds to a vanishing L^2-gradient of I, i.e., $\nabla_{L^2} I(u) = 0$.

2.3.2 Bochner–Sobolev Spaces

For evolutionary partial differential equations we will consider functions $u : [0, T] \to X$ for a time interval $[0, T] \subset \mathbb{R}$. We assume that the Banach space X is separable and say that $u : [0, T] \to X$ is *weakly measurable* if, for all $\varphi \in X'$, the function $t \mapsto \langle \varphi, u(t) \rangle$ is Lebesgue measurable. In this case the *Bochner integral*

$$\int_0^T u(t) \, \mathrm{d}t$$

is well defined with $\left\| \int_0^T u(t) \, \mathrm{d}t \right\|_X \le \int_0^T \|u(t)\|_X \, \mathrm{d}t$. The duality pairing between X and X' will be denoted by $\langle \cdot, \cdot \rangle$.

Definition 2.5 For $1 \le p \le \infty$, the *Bochner space* $L^p([0, T]; X)$ consists of all weakly measurable functions $u : [0, T] \to X$ with $\|u\|_{L^p([0,T];X)} < \infty$, where

$$\|u\|_{L^p([0,T];X)} = \begin{cases} \operatorname{esssup}_{t\in[0,T]}\|u(t)\|_X & \text{if } p=\infty, \\ \Big(\int\limits_{[0,T]} \|u(t)\|^p \,\mathrm{d}t\Big)^{1/p} & \text{if } 1 \le p < \infty. \end{cases}$$

Remark 2.11 The space $L^p([0,T];X)$ is a Banach space when equipped with the norm $\|\cdot\|_{L^p([0,T];X)}$.

Definition 2.6 For $u \in L^1([0,T];X)$ we say that $w \in L^1([0,T];X)$ is the *generalized derivative* of u, denoted by $u' = w$ if for all $\phi \in C_c^\infty((0,T))$, we have

$$\int\limits_0^T \phi'(t)u(t)\,\mathrm{d}t = -\int\limits_0^T \phi(t)w(t)\,\mathrm{d}t.$$

Remark 2.12 Since X is separable one can show that the generalized derivative u' coincides with the weak derivative $\partial_t u$ defined by

$$\int\limits_0^T \langle u, \partial_t\phi\rangle\,\mathrm{d}t = -\int\limits_0^T \langle\partial_t u, \phi\rangle\,\mathrm{d}t$$

for all $\phi \in C^1([0,T];X')$ with $\phi(0) = \phi(T) = 0$.

Definition 2.7 The *Sobolev–Bochner space* $W^{1,p}([0,T];X)$ consists of all functions $u \in L^p([0,T];X)$ with $u' \in L^p([0,T];X)$ and is equipped with the norm

$$\|u\|_{W^{1,p}([0,T];X)} = \begin{cases} \operatorname{esssup}_{t\in[0,T]}(\|u(t)\|_X + \|u'(t)\|_X) & \text{if } p=\infty, \\ \Big(\int\limits_0^T \|u(t)\|_X^p + \|u'(t)\|_X^p \,\mathrm{d}t\Big)^{1/p} & \text{if } 1 \le p < \infty. \end{cases}$$

We write $H^1([0,T];X)$ for $W^{1,2}([0,T];X)$.

Remarks 2.13 (i) We have that $W^{1,p}([0,T];X)$ is a Banach space for $1 \le p \le \infty$. If X is a Hilbert space and $p = 2$, then $W^{1,2}([0,T];X)$ is a Hilbert space denoted by $H^1([0,T];X)$.
(ii) For $1 \le p \le \infty$ and $u \in W^{1,p}([0,T];X)$, we have that $u \in C([0,T];X)$ with $\max_{t\in[0,T]}\|u(t)\| \le c\|u\|_{W^{1,p}([0,T];X)}$.

Definition 2.8 If H is a separable Hilbert space that is identified with its dual H' and such that the inclusion $X \subset H$ is dense and continuous, then (X, H, X') is called a *Gelfand* or an *evolution triple*.

Remark 2.14 For a Gelfand triple (X, H, X') the duality pairing $\langle\varphi, v\rangle$ for $\varphi \in X'$ and $v \in X$ is regarded as a continuous extension of the scalar product on H, i.e., if $\varphi \in X' \cap H'$, then

$$\langle\varphi, v\rangle = (\varphi, v)_H.$$

Below, we always consider an evolution triple (X, H, X'). The Sobolev–Bochner spaces then have the following important properties.

Remarks 2.15 (i) If $u \in L^p([0,T];X)$ with $u' \in L^{p'}([0,T];X')$, then $u \in C([0,T];H)$ with $\max_{t\in[0,T]} \|u(t)\|_H \leq c(\|u\|_{L^p([0,T];X)} + \|u'\|_{L^{p'}([0,T];X')})$ and the *integration-by-parts formula*

$$(u(t_2), v(t_2))_H - (u(t_1), v(t_1))_H = \int\limits_{t_1}^{t_2} \langle u'(t), v(t)\rangle + \langle v'(t), u(t)\rangle \,\mathrm{d}t$$

holds for all $v \in L^p([0,T];X)$ with $v' \in L^{p'}([0,T];X)$ and $t_1, t_2 \in [0,T]$. In particular, we have

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|_H^2 = \langle u'(t), u(t)\rangle$$

for almost every $t \in [0,T]$.
(ii) If X is compactly embedded in H, $1 < p < \infty$, and $1 < q \leq \infty$, then according to the *Aubin–Lions lemma* the inclusion $L^p([0,T];X) \cap W^{1,q}([0,T];X') \subset L^p([0,T];H)$ is compact.
(iii) For $1 \leq p < \infty$ the space $L^p([0,T];X)$ is separable. In particular, if $(f_n)_{n\in\mathbb{N}} \subset L^p(I)$ and $(v_n)_{n\in\mathbb{N}} \subset X$ are dense subsets, then $\mathrm{span}\{f_n v_m : n, m \in \mathbb{N}\}$ is dense in $L^p(I;X)$.
(iv) If $g \in L^{p'}([0,T];X')$, then the mapping $f \mapsto \int_0^T \langle f(t), g(t)\rangle \,\mathrm{d}t$, defined for every $f \in L^p([0,T];X)$, belongs to $(L^p([0,T];X))'$ for $1 \leq p \leq \infty$. If $1 < p < \infty$, we have that $L^p([0,T];X)$ is reflexive provided that X is reflexive. In particular, for $1 \leq p < \infty$ we have $(L^p([0,T];X))' = L^{p'}([0,T];X')$.
(v) We have that $L^2(I;H)$ is a Hilbert space.

2.3.3 Existence Theory for Gradient Flows

We consider a Fréchet-differentiable functional $I : X \to \mathbb{R}$ with $DI : X \to X'$ and we want to derive conditions that guarantee existence of solutions for the H-gradient flow of I formally defined by

$$\partial_t u = -\nabla_H I(u), \quad u(0) = u_0.$$

We always let (X, H, X') be an evolution triple and assume that an abstract Poincaré inequality holds, i.e., that for a seminorm $|\cdot|_X$ on X we have

$$\|u\|_X \leq c_P(|u|_X + \|u\|_H).$$

for all $u \in X$.

Definition 2.9 Given $u_0 \in H$, we say that $u \in L^p([0,T];X)$ is a *solution of the H-gradient flow for I* if $u' \in L^{p'}([0,T];X')$ and for almost every $t \in [0,T]$ and every $v \in X$, we have that

$$\langle u'(t), v\rangle + DI(u)[v] = 0$$

and $u(0) = u_0$.

Example 2.3 For $I(u) = (1/2)\int_\Omega |\nabla u|^2 \,\mathrm{d}x$ defined on $H_0^1(\Omega)$, the $L^2(\Omega)$-gradient flow is the linear heat equation $\partial_t u - \Delta u = 0$.

We follow the *Rothe method* to construct solutions. This method consists of three steps: First, we consider an implicit time discretization that replaces the time derivative by difference quotients and establishes the existence of approximations. In the second step, a priori bounds that allow us to extract weakly convergent subsequences of the approximations as the time-step size tends to zero are proved. Finally, we pass to the limit and try to show that weak limits are solutions of the gradient flow.

Definition 2.10 The functional $I : X \to \mathbb{R}$ is called *semicoercive* if there exist $s > 0$, $c_1 > 0$, and $c_2 \in \mathbb{R}$ such that

$$I(v) \ge c_1 |v|_X^s - c_2 \|v\|_H^2$$

for all $v \in X$.

Proposition 2.1 (Implicit Euler scheme) *Assume that I is semicoercive and weakly lower semicontinuous. Then for every $\tau > 0$ with $4\tau c_2 < 1$ and $k = 1, 2, \dots, K$, $K = \lceil T/\tau \rceil$, the functionals $I^k : X \to \mathbb{R}$,*

$$u \mapsto I^k(u) = \frac{1}{2\tau}\|u - u^{k-1}\|_H^2 + I(u),$$

with $u^0 = u_0$ have minimizers that satisfy

$$(d_t u^k, v)_H + DI(u^k)[v] = 0$$

for all $v \in X$ with the backward difference quotient $d_t u^k = (u^k - u^{k-1})/\tau$.

Proof Since I^k is coercive, bounded from below, and weakly lower semicontinuous, the direct method in the calculus of variations implies the existence of a minimum. Since I and $v \mapsto \|v\|_H^2$ are Fréchet-differentiable, the minimizers satisfy the asserted equations. □

Remarks 2.16 (i) More generally, one can consider a pseudomonotone operator $A : X \to X'$ and look for a solution u^k of the equation $(d_t u^k, v)_H + A(u^k)[v] = 0$ for all $v \in X$.
(ii) We have that u^k is uniquely defined if I^k is strictly convex. This is often satisfied for τ sufficiently small, i.e., if I is semiconvex.

For the proof of the a priori bounds two important ingredients are required. The first is based on the binomial formula $2(a-b)a = (a-b)^2 + (a^2 - b^2)$ and shows that

$$\frac{1}{\tau}(u^k - u^{k-1}, u^k)_H = \frac{1}{2\tau}\|u^k - u^{k-1}\|_H^2 + \frac{1}{2\tau}\big(\|u^k\|_H^2 - \|u^{k-1}\|_H^2\big)$$

for $k = 1, 2, \ldots, K$. Equivalently, we have

$$(d_t u^k, u^k)_H = \frac{\tau}{2}\|d_t u^k\|_H^2 + \frac{1}{2} d_t \|u^k\|_H^2$$

which is a discrete version of the identity $2\langle u', u\rangle = (d/dt)\|u\|_H^2$. The second ingredient is the following discrete Gronwall lemma.

Lemma 2.2 (Discrete Gronwall lemma) *Let* $(y^\ell)_{\ell=0,1,\ldots,L}$ *be a sequence of nonnegative real numbers such that for nonnegative real numbers* $a_0, b_0, b_1, \ldots, b_{L-1}$ *and* $\ell = 0, 1, \ldots, L$*, we have*

$$y^\ell \le a_0 + \sum_{k=0}^{\ell-1} b_k y^k.$$

Then we have $\max_{\ell=0,1,\ldots,L} y^\ell \le a_0 \exp\big(\sum_{k=0}^{\ell-1} b_k\big)$.

Proof The proof follows from an inductive argument. □

We also have to assume a coerciveness property for the mapping $DI : X \to X'$.

Definition 2.11 We say that $DI : X \to X'$ is *semicoercive and bounded* if there exist $p \in (1, \infty)$, $c_1' > 0$, $c_2' \in \mathbb{R}$, and $c_3' > 0$ such that

$$DI(v)[v] \ge c_1' |v|_X^p - c_2' \|v\|_H^2$$

and

$$\|DI(v)\|_{X'} \le c_3'(1 + \|v\|_X^{p-1})$$

for all $v \in X$.

Proposition 2.2 (A priori bounds) *Suppose that* $DI : X \to X'$ *is semicoercive and bounded. If* $4\tau c_2' \le 1$*, then we have*

$$\max_{\ell=0,1,\ldots,K} \|u^\ell\|_H + \tau \sum_{k=1}^K \|u^k\|_X^p + \tau \sum_{k=1}^K \|d_t u^k\|_{X'}^{p'} + \tau \sum_{k=1}^K \|DI(u^k)\|_{X'}^{p'} \le C_0$$

with a constant $C_0 > 0$ *that depends on* p*,* T*,* u_0*,* c_P*,* c_1'*,* c_2'*, and* c_3'*.*

Proof Since $(d_t u^k, v)_H + DI(u^k)[v] = 0$ for all $v \in X$, we obtain by choosing $v = u^k$ that

$$\frac{\tau}{2}\|d_t u^k\|_H^2 + \frac{1}{2} d_t \|u^k\|_H^2 + c_1' |u^k|_X^p - c_2' \|u^k\|_H^2 \le (d_t u^k, u^k)_H + DI(u^k)[u^k] = 0.$$

Multiplication by τ and summation over $k = 1, 2, \dots, \ell$ for $1 \le \ell \le K$ lead to

$$\frac{1}{2}\|u^\ell\|_H^2 + \tau \sum_{k=1}^{\ell} \frac{\tau}{2}\|d_t u^k\|_H^2 + c_1' \tau \sum_{k=1}^{\ell} |u^k|_X^p \le \frac{1}{2}\|u^0\|_H^2 + c_2' \tau \sum_{k=1}^{\ell} \|u^k\|_H^2,$$

where we used the telescope effect $\tau \sum_{k=1}^{\ell} d_t \|u^k\|_H^2 = \|u^\ell\|_H^2 - \|u^0\|_H^2$. Since the second term on the left-hand side is nonnegative and since $4c_2'\tau \le 1$ so that we can absorb $c_2'\tau\|u^\ell\|_H^2$ on the left-hand side, we find that

$$\frac{1}{4}\|u^\ell\|_H^2 + c_1' \tau \sum_{k=1}^{\ell} |u^k|_X^p \le \frac{1}{2}\|u^0\|_H^2 + c_2' \tau \sum_{k=1}^{\ell-1} \|u^k\|_H^2.$$

For $\ell = 0, 1, \dots, K$, we set

$$y^\ell = \frac{1}{4}\|u^\ell\|_H^2 + c_1' \tau \sum_{k=1}^{\ell} |u^k|_X^p,$$

$a_0 = (1/2)\|u^0\|_H^2$ and $b = 4c_2'\tau$ so that

$$y^\ell \le a_0 + \sum_{k=1}^{\ell-1} b y^k$$

and we are in the situation to apply the discrete Gronwall lemma. This shows that

$$\max_{\ell=0,\dots,K} \frac{1}{4}\|u^\ell\|_H^2 + c_1' \tau \sum_{\ell=1}^{K} |u^k|_X^p \le \frac{1}{2}\|u^0\|_H^2 \exp\big(4c_2' \sum_{k=1}^{K-1} \tau\big) \le \frac{1}{2}\|u^0\|_H^2 \exp\big(4c_2' T\big)$$

and proves the bound for the first term on the left-hand side. The second bound follows with the abstract Poincaré inequality $\|u\|_X \le c_P(|u|_X + \|u\|_H)$ and

$$\|u^k\|_X^p \le c_P^p (|u^k|_X + \|u^k\|_H)^p \le c_P^p 2^{p-1}\big(|u^k|_X^p + c_P^p \|u^k\|_H^p\big).$$

For the third and fourth term on the left-hand side of the bound, we note that with the boundedness of DI and $(p-1)p' = p$, it follows that

$$\tau \sum_{k=1}^{L} \|DI(u^k)\|_{X'}^{p'} \le \tau (c_3')^{p'} \sum_{k=1}^{K} (1 + \|u^k\|_X^{p-1})^{p'} \le \tau (c_3')^{p'} 2^{p'-1} \sum_{k=1}^{K} (1 + \|u^k\|_X^p),$$

and the right-hand side is bounded according to the previous bounds. This proves the fourth estimate and the third bound follows immediately since $\|d_t u^k\|_{X'} = \|DI(u^k)\|_{X'}$ due to the identity $(d_t u^k, v) = -DI(u^k)[v]$ for all $v \in X$. □

To identify the limits of the approximations we define interpolants of the approximations $(u^k)_{k=0,\dots,K}$.

Definition 2.12 Given a time-step size $\tau > 0$ and a sequence $(u^k)_{k=0,\dots,K} \subset H$ for $K = \lceil T/\tau \rceil$, we set $t_k = k\tau$ for $k = 0, 1, \dots, K$ and define the piecewise constant and piecewise affine interpolants $u_\tau{}^-, u_\tau^+, \widehat{u}_\tau : [0, T] \to H$ for $t \in (t_{k-1}, t_k)$ by

$$u_\tau^-(t) = u^{k-1}, \quad u_\tau^+(t) = u^k, \quad \widehat{u}_\tau(t) = \frac{t - t_{k-1}}{\tau} u^k + \frac{t_k - t}{\tau} u^{k-1}.$$

The construction of the interpolants is illustrated in Fig. 2.14.

Remarks 2.17 (i) We have $\widehat{u}_\tau \in W^{1,\infty}([0, T]; H)$ with $\widehat{u}_\tau' = d_t u^k$ on (t_{k-1}, t_k) for $k = 1, 2, \dots, K$. Moreover, $u_\tau^+, u_\tau^- \in L^\infty([0, T]; H)$ and, e.g.,

$$\|u_\tau^+\|_{L^p([0,T];X)}^p \le \tau \sum_{k=1}^{K} \|u^k\|_X^p$$

with equality if $K\tau = T$.
(ii) We have $u_\tau^+(t) - u_\tau^-(t) = \tau \widehat{u}_\tau'(t)$ and $\widehat{u}_\tau(t) = u_\tau^-(t) + (t - t_{k-1})\widehat{u}_\tau'(t) = u_\tau^+(t) - (t_k - t)\widehat{u}_\tau'(t)$ for almost every $t \in (t_{k-1}, t_k)$.
(iii) If $\|\widehat{u}_\tau'\|_{L^1([0,T];X')} \le c$ for all $\tau > 0$, then it follows that $u_\tau^+ - u_\tau^- \to 0$ and $\widehat{u}_\tau - u_\tau^\pm \to 0$ in $L^1([0, T]; X')$ as $\tau \to 0$. In particular, all interpolants have the same limit if these exists.

Lemma 2.3 (Discrete evolution equation) *With the interpolants of the approximations* $(u^k)_{k=0,\dots,K}$, *we have*

$$(\widehat{u}_\tau'(t), v)_H + DI(u_\tau^+(t))[v] = 0$$

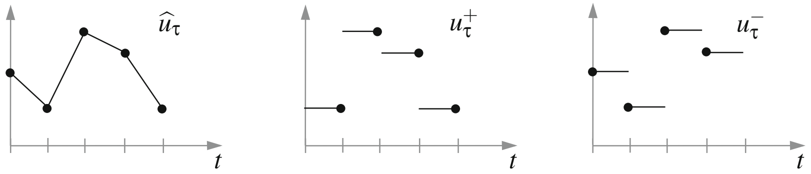


Fig. 2.14 Continuous interpolant $\widehat{u}_\tau$ (*left*) and piecewise constant interpolants u_τ^+ (*middle*) and u_τ^- (*right*)

for all $v \in X$ and almost every $t \in [0, T]$. Moreover, we have for all $\tau > 0$

$$\|u_\tau^+\|_{L^\infty([0,T];H)} + \|u_\tau^+\|_{L^p([0,T];X)} + \|\widehat{u}_\tau\|_{W^{1,p'}([0,T];X')} + \|DI(u_\tau^+)\|_{L^{p'}([0,T];X')} \le C_0.$$

Proof The identity follows directly from $(d_t u^k, v)_H + DI(u^k)[v] = 0$ for $k = 1, 2, \dots, K$ and all $v \in X$ with the definitions of the interpolants $\widehat{u}_\tau$ and u_τ^+. With the triangle inequality and $|t - t_k| \le \tau$ for $t \in (t_{k-1}, t_k)$, we observe that

$$\|\widehat{u}_\tau\|_{L^{p'}([0,T];X')} \le c_P T^{1/p'} \|u^+\|_{L^\infty([0,T];H)} + \tau \|\widehat{u}'_\tau\|_{L^{p'}([0,T];X')}.$$

The a priori bounds of Proposition 2.2 together with, e.g.,

$$\tau \sum_{k=1}^{K} \|u^k\|_X^p \ge \int_0^T \|u_\tau^+\|_X^p \, \mathrm{d}t,$$

where we used $K\tau \ge T$, imply the a priori bounds. □

The bounds for the interpolants allow us to select accumulation points.

Proposition 2.3 (Selection of a limit) *Assume that X is compactly embedded in H. Then there exist $u \in L^p([0, T]; X) \cap W^{1,p'}([0, T]; X')$ and $\xi \in L^{p'}([0, T]; X')$ such that for a sequence $(\tau_n)_{n\in\mathbb{N}}$ of positive numbers with $\tau_n \to 0$ as $n \to \infty$, we have*

$$\begin{aligned} \widehat{u}_{\tau_n}, u_{\tau_n} &\rightharpoonup^* u \quad \text{in } L^\infty([0, T]; H), \\ \widehat{u}_{\tau_n}, u_{\tau_n} &\rightharpoonup u \quad \text{in } L^p([0, T]; X), \\ \widehat{u}_{\tau_n} &\rightharpoonup u \quad \text{in } W^{1,p'}([0, T]; X'), \\ DI(u_{\tau_n}^+) &\rightharpoonup \xi \quad \text{in } L^{p'}([0, T]; X'). \end{aligned}$$

We have $u \in C([0, T]; H)$ with $u(0) = u_0$ and

$$\langle u'(t), v\rangle + \langle \xi(t), v\rangle = 0$$

for almost every $t \in [0, T]$ and all $v \in X$. In particular, if $\xi = DI(u)$, then u is a solution of the H-gradient flow for I.

Proof For a sequence $(\tau_n)_{n\in\mathbb{N}}$ of positive numbers with $\tau_n \to 0$ as $n \to \infty$, the a priori bounds yield the existence of weak limits for an appropriate subsequence which is not relabeled. Due to the bound for $\widehat{u}'_\tau$, the weak limits coincide. Multiplying the discrete evolution equation of Lemma 2.3 by $\phi \in C([0, T])$ and integrating the resulting identity over $[0, T]$ we find that

$$\int_0^T \langle \widehat{u}'_{\tau_n}, \phi v\rangle + DI(u^+_{\tau_n})[\phi v]\,\mathrm{d}t = 0$$

for every $v \in X$. Since $\phi v \in L^p([0,T];X)$ we can pass to the limit $n \to \infty$ in the equation and obtain

$$\int_0^T \langle u', \phi v\rangle + \langle \xi, \phi v\rangle\,\mathrm{d}t = 0.$$

Since this holds for every $\phi \in C([0,T])$ we deduce the asserted equation. The mapping $v \mapsto v(0)$ defines a bounded linear operator $L^p([0,T];X) \cap W^{1,p'}([0,T];X') \to H$ which is weakly continuous. Since $\widehat{u}_{\tau_n}(0) = u_0$ for all $n \in \mathbb{N}$, we deduce that $u(0) = u_0$. By continuous embeddings we also have $u \in C([0,T];H)$ which implies the continuous attainment of the initial data. □

Remark 2.18 The assumed identity $\xi = DI(u)$ in $L^{p'}([0,T];X')$, i.e., the convergence $DI(u^+_{\tau_n}) \rightharpoonup DI(u)$ can in general only be established under additional conditions on DI and requires special techniques from nonlinear functional analysis, e.g., based on concepts of pseudomonotonicity.

Example 2.4 For $F \in C^1(\mathbb{R})$ with $0 \le F(s) \le c_F(1+|s|^2)$ and $f(s) = F'(s)$ such that $|f(s)| \le c'_F(1+|s|)$, we consider

$$I(u) = \frac{1}{2}\int_\Omega |\nabla u|^2\,\mathrm{d}x + \int_\Omega F(u)\,\mathrm{d}x.$$

Then, for $X = H^1_0(\Omega)$ and $H = L^2(\Omega)$, the conditions of the previous propositions are satisfied with $p = 2$ and

$$DI(u)[v] = \int_\Omega \nabla u \cdot \nabla v\,\mathrm{d}x + \int_\Omega f(u)v\,\mathrm{d}x.$$

We have $u^+_{\tau_n} \rightharpoonup u \in L^2([0,T];H^1_0(\Omega))$ so that $\nabla u^+_{\tau_n} \rightharpoonup \nabla u$ in $L^2([0,T];L^2(\Omega;\mathbb{R}^d))$ and thus

$$\int_0^T\int_\Omega \nabla u^+_{\tau_n} \cdot \nabla w\,\mathrm{d}x\,\mathrm{d}t \to \int_0^T\int_\Omega \nabla u \cdot \nabla w\,\mathrm{d}x\,\mathrm{d}t$$

for all $w \in L^2([0,T];H^1_0(\Omega))$. The compactness of the embedding

$$L^2([0,T];H^1_0(\Omega)) \cap W^{1,2}([0,T];H^1_0(\Omega)') \to L^2([0,T];L^2(\Omega)) = L^2([0,T]\times\Omega)$$

in combination with the generalized dominated convergence theorem shows that

$$\int_0^T \int_\Omega f(u_{\tau_n}^+)w \,\mathrm{d}x \,\mathrm{d}t \to \int_0^T \int_\Omega f(u)w \,\mathrm{d}x \,\mathrm{d}t.$$

Altogether this proves that

$$\int_0^T DI(u_{\tau_n}^+)[w] \,\mathrm{d}t \to \int_0^T \int_\Omega \nabla u \cdot \nabla w \,\mathrm{d}x \,\mathrm{d}t + \int_0^T \int_\Omega f(u)w \,\mathrm{d}x \,\mathrm{d}t,$$

i.e., $\xi = DI(u)$.

Remarks 2.19 (i) For the semilinear heat equation $\partial_t u = \Delta u - f(u)$ of Example 2.4, one can establish the existence of a solution under more general conditions on f. Moreover, one can prove stronger a priori bounds and the energy law

$$I(u(T')) + \int_0^{T'} \|u'(t)\|_{L^2(\Omega)}^2 \,\mathrm{d}t \le I(u_0)$$

for almost every $T' \in [0, T]$ provided $u_0 \in H_0^1(\Omega)$. The key ingredient is the convexity of I in the highest-order term.
(ii) An alternative method to establish the existence of solutions for gradient flows is the Galerkin method which is based on a discretization in space. This leads to a sequence of ordinary differential equations on finite-dimensional spaces and with appropriate a priori bounds, one can then show under appropriate conditions that the approximate solutions converge to a solution as the dimension tends to infinity.

2.3.4 Subdifferential Flows

The estimates for discretized gradient flows can be significantly improved if the functional I is convex, since then we would have

$$I(u^k) + DI(u^k)[u^{k-1} - u^k] \le I(u^{k-1}).$$

In particular, choosing $v = d_t u^k$ in the identity $(d_t u^k, v)_H + DI(u^k)[v] = 0$ gives

$$\tau \|d_t u^k\|_H^2 + I(u^k) \le I(u^{k-1})$$

and a summation over k yields the a priori bound $I(u^L) + \tau \sum_{k=1}^L \|d_t u^k\|_H^2 \le I(u^0)$. With these observations it is possible to establish a theory for convex functionals that are not differentiable. We always consider a Hilbert space H that is identified with its dual.

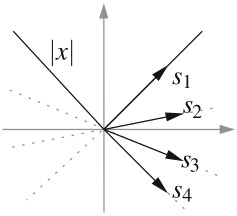


Fig. 2.15 Subdifferential of the function $x \mapsto |x|$ at $x = 0$; the arrows s_1, s_2, s_3, s_4 indicate subgradients at 0 which are the slopes of supporting hyperplanes at 0

Definition 2.13 We say that a functional $I : H \to \mathbb{R} \cup \{+\infty\}$ belongs to the class $\Gamma(H)$ if it is convex, lower semicontinuous, i.e., $I(u) \le \liminf_{n\to\infty} I(u_n)$ whenever $u_n \to u$ in H as $n \to \infty$, and *proper*, i.e., there exists $u \in H$ with $I(u) \in \mathbb{R}$.

We assume that $I \in \Gamma(H)$ below.

Definition 2.14 The *subdifferential* $\partial I : H \to 2^H$ of I associates to every $u \in H$ the set

$$\partial I(u) = \{v \in H : I(w) \ge I(u) + (v, w - u)_H \text{ for all } w \in H\}.$$

The elements in $\partial I(u)$ are called *subgradients* of I at u.

Example 2.5 For $F(x) = |x|$, $x \in \mathbb{R}$, we have $\partial F(0) = [-1, 1]$, cf. Fig. 2.15.

Remarks 2.20 (i) The subdifferential $\partial I(u)$ consists of all slopes of affine functions that are below I and that intersect the graph of I at u.
(ii) For all $u_1, u_2 \in H$ and $v_1 \in \partial I(u_1)$, $v_2 \in \partial I(u_2)$ we have the *monotonicity estimate*

$$(v_1 - v_2, u_1 - u_2)_H \ge 0.$$

(iii) We have $0 \in \partial I(u)$ if and only if $u \in H$ is a global minimum for I.
(iv) We have $\partial I(u) = \{s\}$ for $s \in H$ if and only if I is Gâteaux-differentiable at u.
(v) For $I, J \in \Gamma(H)$ we have $\partial(I + J) \subset \partial I + \partial J$, and if there exists a point at which I and J are finite and I or J is continuous, we have equality.

Theorem 2.7 (Resolvent operator) *Let $I \in \Gamma(H)$. For every $w \in H$ and $\lambda > 0$ there exists a unique $u \in H$ with*

$$u + \lambda \partial I(u) \ni w.$$

This defines the resolvent operator $u = R_\lambda(w) = (\mathrm{Id} + \lambda \partial I)^{-1}(w)$.

Proof For a short proof we make the simplifying but nonrestrictive assumption that

$$I(v) \ge -c_1 - c_2 \|v\|_H.$$

For $\lambda > 0$ and $w \in H$ we consider the minimization problem defined through the functional

$$I_{\lambda,w}(u) = \frac{1}{2\lambda}\|u-w\|_H^2 + I(u) = \frac{1}{2\lambda}\|u\|_H^2 - \frac{1}{\lambda}(u,w)_H + \frac{1}{2\lambda}\|w\|_H^2 + I(u).$$

The identity $2(a^2+b^2) = (a+b)^2 + (a-b)^2$ and the convexity of I show that for $u_1, u_2 \in H$ we have

$$\begin{aligned}&\frac{1}{2}I_{\lambda,w}(u_1) + \frac{1}{2}I_{\lambda,w}(u_2) - I_{\lambda,w}\big(\frac{u_1+u_2}{2}\big)\\ &\qquad = \frac{1}{8\lambda}\|u_1-u_2\|_H^2 + \frac{1}{2}I(u_1) + \frac{1}{2}I(u_2) - I\big(\frac{u_1+u_2}{2}\big) \geq \frac{1}{8\lambda}\|u_1-u_2\|_H^2,\end{aligned}$$

i.e., $I_{\lambda,w}$ is *strictly convex*. Thus, if $I_{\lambda,w}$ has a minimizer, then it is unique. Moreover, $u \in H$ minimizes $I_{\lambda,w}$ if and only if $0 \in \partial I_{\lambda,w}(u) = (1/\lambda)(u-w)+\partial I(u)$. It remains to show that there exists a minimizer. Since $I_{\lambda,w}$ is convex and lower semicontinuous it follows that I is weakly lower semicontinuous. We also have that $I_{\lambda,w}$ is coercive since two applications of Young's inequality lead to

$$\begin{aligned}I_{\lambda,w}(v) &\geq \frac{1}{2\lambda}\|v\|_H^2 - \frac{1}{\lambda}(v,w)_H + \frac{1}{2\lambda}\|w\|_H^2 - c_1 - c_2\|v\|_H\\ &\geq \frac{1}{4\lambda}\|v\|_H^2 - \frac{4}{\lambda}\|w\|_H^2 - c_1 - 4\lambda c_2^2.\end{aligned}$$

This estimate also proves the boundedness from below. The direct method in the calculus of variations thus implies the existence of a minimizer. □

Definition 2.15 The *Yosida regularization* $A_\lambda : H \to H$ is for $w \in H$ defined by $A_\lambda(w) = (1/\lambda)(w - R_\lambda(w))$.

Remark 2.21 The resolvent operator satisfies $\lim_{\lambda\to 0} R_\lambda(w) = w$. We have that A_λ is Lipschitz continuous with Lipschitz constant $2/\lambda$ and approximates ∂I in the sense that $A_\lambda(w) \in \partial I(R_\lambda w)$.

The theorem about the resolvent operator implies that for a time-step size $\tau > 0$ and an initial $u^0 \in H$, there exists a unique sequence $(u^k)_{k=0,\dots,L} \subset H$ with

$$d_t u^k \in -\partial I(u^k)$$

since this is equivalent to $u^k = R_\tau(u^{k-1})$. We expect that as $\tau \to 0$ the approximations converge to a solution of the *subdifferential flow*

$$u' \in -\partial I(u), \quad u(0) = u_0.$$

Related a priori bounds that permit a corresponding passage to a limit will be discussed in Chap. 4.

Theorem 2.8 (Subdifferential flow, [2]) *For every $u_0 \in H$ such that $\partial I(u_0) \neq \emptyset$ and every $T > 0$, there exists a unique function $u \in C([0,T];H)$ with $u' \in L^\infty([0,T];H)$ such that $u(0) = u_0$, $\partial I(u(t)) \neq \emptyset$ for every $t \in [0,T]$, and*

$$u'(t) \in -\partial I(u(t))$$

for almost every $t \in [0,T]$.

Proof The existence of a solution is established by considering for every $\lambda > 0$ the problem

$$\partial_t u_\lambda = -A_\lambda(u_\lambda), \quad u_\lambda(0) = u_0$$

and studying the limit $\lambda \to 0$. Uniqueness of solutions follows from the convexity of I, i.e., if u_1 and u_2 are solutions then the monotonicity property of I shows that

$$-(u_1'(t) - u_2'(t), u_1(t) - u_2(t))_H \geq 0$$

for almost every $t \in [0,T]$ and this implies that

$$\frac{1}{2}\frac{d}{dt}\|(u_1 - u_2)(t)\|_H^2 = (u_1'(t) - u_2'(t), u_1(t) - u_2(t))_H \leq 0.$$

Since $u_1(0) = u_2(0)$ we deduce that $u_1(t) = u_2(t)$ for every $t \in [0,T]$. □

Remarks 2.22 (i) Negative subgradients are in general no descent directions. For the subdifferential flow one can however show that $u'(t) = -\partial^0 I(u(t))$ for almost every $t \in [0,T]$, where $\partial^0 I(v)$ is the subgradient $s \in \partial I(v)$ with minimal norm, i.e., $\|\partial^0 I(v)\|_H = \min_{r \in \partial I(v)} \|r\|_H$.
(ii) If $\partial I(u_0) = \emptyset$, then there exists a unique solution $u \in C([0,T];H)$ such that $t^{1/2}u' \in L^2([0,T];H)$.

References

1. Attouch, H., Buttazzo, G., Michaille, G.: Variational Analysis in Sobolev and BV Spaces. MPS/SIAM Series on Optimization, vol. 6. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2006)
2. Brézis, H.R.: Opérateurs Maximaux Monotones et Semi-groupes de Contractions dans les Espaces de Hilbert. North-Holland Publishing Co., Amsterdam (1973). North-Holland Mathematics Studies, No. 5. Notas de Matemática (50)
3. Ciarlet, P.G.: Mathematical Elasticity. Vol. I: Three-Dimensional Elasticity. Studies in Mathematics and its Applications, vol. 20. North-Holland Publishing Co., Amsterdam (1988)
4. Dacorogna, B.: Direct Methods in the Calculus of Variations. Applied Mathematical Sciences, vol. 78, 2nd edn. Springer, New York (2008)
5. Eck, C., Garcke, H., Knabner, P.: Mathematische Modellierung, 2nd edn. Springer Lehrbuch. Springer, Berlin (2011)
6. Evans, L.C.: Partial Differential Equations. Graduate Studies in Mathematics, vol. 19, 2nd edn. American Mathematical Society, Providence (2010)

7. Roubíček, T.: Nonlinear Partial Differential Equations with Applications. International Series of Numerical Mathematics, vol. 153, 2nd edn. Birkhäuser/Springer Basel AG, Basel (2013)
8. Ružička, M.: Nichtlineare Funktional Analysis. Springer, Berlin-Heidelberg-New York (2004)
9. Salsa, S., Vegni, F.M.G., Zaretti, A., Zunino, P.: A Primer on PDEs. Unitext, vol. 65, italian edn. Springer, Milan (2013)
10. Struwe, M.: Variational Methods, 4th edn. Springer, Berlin (2008)
11. Temam, R., Miranville, A.: Mathematical Modeling in Continuum Mechanics, 2nd edn. Cambridge University Press, Cambridge (2005)

Chapter 3
FEM for Linear Problems

3.1 Interpolation with Finite Elements

We review in this section the basic framework for analyzing finite element methods. We refer the reader to the textbooks and lecture notes [2, 3, 5, 6, 8–11] for further details. A review of the historical development of the finite element method can be found in [7].

3.1.1 Abstract Finite Elements

For a set $T \subset \mathbb{R}^d$ and an integer $k \geq 0$, we let $P_k(T)$ denote the set of polynomials of total degree less than or equal to k restricted to T.

Definition 3.1 A *finite element* is a triple (T, P_T, K_T) consisting of a closed set $T \subset \mathbb{R}^d$, a space of polynomials P_T with $\dim P_T = R$, and a set $K_T = \{\chi_1, \ldots, \chi_R\}$ of linear functionals on $C^\infty(T)$ such that (a) if for $q \in P_T$ we have $\chi(q) = 0$ for all $\chi \in K_T$, then $q = 0$, (b) there exists $m \geq 1$ with $P_{m-1}(T) \subset P_T$, and (c) there exists $p \in [1, \infty]$ such that every $\chi \in K_T$ extends to a bounded linear operator on $W^{m,p}(T)$.

Definition 3.2 Given a finite element (T, P_T, K_T) and $v \in W^{m,p}(T)$ the *interpolant* $I_T v \in P_T$ is the uniquely defined function in P_T that satisfies $\chi(I_T v) = \chi(v)$ for all $\chi \in K_T$.

Example 3.1 For a line segment, triangle, or tetrahedron $T = \mathrm{conv}\{z_0, z_1, \ldots, z_d\} \subset \mathbb{R}^d$, $d = 1, 2, 3$, respectively, set $P_T = P_1(T)$, $K_T = \{\chi_0, \chi_1, \ldots, \chi_d\}$ with $\chi_j(v) = v(z_j)$ for $j = 0, 1, \ldots, d$ and $v \in C^\infty(T)$. Then (T, P_T, K_T) is a finite element with $m = 2$ and $p = 2$ called a $P1$ element.

S. Bartels, *Numerical Methods for Nonlinear Partial Differential Equations*,
Springer Series in Computational Mathematics 47,
DOI 10.1007/978-3-319-13797-1_3

The properties of interpolants can be analyzed with the Bramble–Hilbert lemma.

Theorem 3.1 (Bramble–Hilbert lemma) *Let* $1 \le p < \infty$ *and let* $F : W^{m,p}(T) \to \mathbb{R}$ *be a bounded and quasisublinear functional, i.e., there exist* $c_1, c_2 > 0$ *such that for all* $v, w \in W^{m,p}(T)$*, we have* $|F(v)| \le c_1 \|v\|_{W^{m,p}(T)}$ *and* $|F(v+w)| \le c_2(|F(v)| + |F(w)|)$ *and assume that* F *vanishes on* $P_{m-1}(T)$*. Then there exists* $c_0 > 0$ *such that*

$$|F(v)| \le c_0 c_1 c_2 |D^m v|_{L^p(T)}$$

for all $v \in W^{m,p}(T)$*.*

Proof Let $v \in W^{m,p}(T)$. For all $q \in P_{m-1}(T)$ we have that

$$|F(v)| \le c_2 |F(v-q)| \le c_1 c_2 \|v-q\|_{W^{m,p}(T)}.$$

There exists a uniquely defined $q \in P_{m-1}(T)$ satisfying $\int_T D^\alpha(v-q)\,\mathrm{d}x = 0$ for all $\alpha \in \mathbb{N}^d$ with $|\alpha| < m$ and a generalized Poincaré inequality implies the estimate $\|v-q\|_{W^{m,p}(T)} \le c_0 \|D^m(v-q)\|_{L^p(T)}$. Since $D^m q = 0$ we deduce the assertion. □

Corollary 3.1 (Interpolation stability) *Let* (T, P_T, K_T) *be a finite element and* $|\cdot|_S$ *a seminorm on* $W^{m,p}(T)$ *with* $|v|_S \le c_S \|v\|_{W^{m,p}(T)}$ *for all* $v \in W^{m,p}(T)$*. Then we have*

$$|v - I_T v|_S \le c_S \|D^m v\|_{L^p(T)}$$

for all $v \in W^{m,p}(T)$*.*

Proof We define $F(v) = |v - I_T v|_S$ and note that F is sublinear. There exists a uniquely defined dual basis $(\psi_1, \dots, \psi_R) \subset P_T$ with $\chi_j(\psi_k) = \delta_{jk}$ for $1 \le j, k \le R$. We then have $I_T(v) = \sum_{j=1}^R \chi_j(v)\psi_j$ and using $|\chi_j(v)| \le c_b \|v\|_{W^{m,p}(T)}$ for all $v \in W^{m,p}(T)$ and $j = 1, \dots, R$, it follows that

$$|F(v)| \le |v|_S + |I_T v|_S \le (c_S + c_b \max_{j=1,\dots,R} |\psi_j|_S)\|v\|_{W^{m,p}(T)},$$

i.e., F is bounded. Obviously, $F(q) = 0$ for all $q \in P_T$, and hence the conditions of the Bramble–Hilbert lemma are satisfied. □

3.1.2 P1 Finite Elements

We consider a bounded, polyhedral Lipschitz domain $\Omega \subset \mathbb{R}^d$ and a given partition $\Gamma_\mathrm{D} \cup \Gamma_\mathrm{N} = \partial\Omega$.

Definition 3.3 A (*conforming*) *triangulation* $\mathcal{T}_h$ of Ω is a set $\mathcal{T}_h = \{T_1, T_2, \dots, T_L\}$ of closed intervals, triangles, or tetrahedra for $d = 1, 2, 3$, respectively, called *elements*, such that $\overline{\Omega} = \cup_{T \in \mathcal{T}_h} T$ and the intersection of distinct $T_1, T_2 \in \mathcal{T}_h$ is either

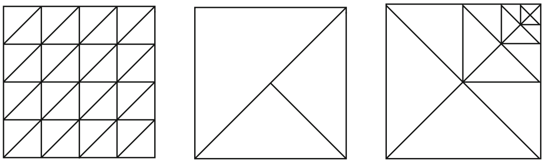
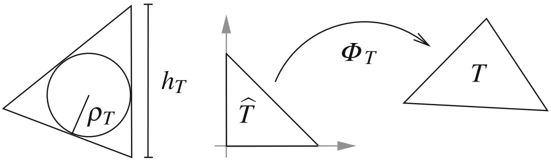


Fig. 3.1 Uniform conforming triangulation, nonconforming triangulation with a hanging node, and locally refined conforming triangulation (from *left* to *right*)

Fig. 3.2 Diameter h_T and inner radius ρ_T of a *triangle* T (*left*); affine transformation from a reference element (*right*)



empty or an entire subsimplex, and the sets Γ_{D} and $\overline{\Gamma}_{\mathrm{N}}$ are matched exactly by the union of sides of elements, cf. Fig. 3.1.

For an element $T \in \mathcal{T}_h$ we set $h_T = \operatorname{diam}(T)$ and let ρ_T denote the diameter of the largest ball contained in T, cf. Fig. 3.2. The importance of the Bramble–Hilbert lemma lies in the scaling properties of the seminorm in $W^{m,p}(T)$ with respect to affine transformations.

Proposition 3.1 (Affine transformations) *Let* $\widehat{T} = \operatorname{conv}\{\widehat{z}_0, \widehat{z}_1, \dots, \widehat{z}_d\}$, *where* $\widehat{z}_0 = 0$ *and* $\widehat{z}_j = e_j$ *with canonical basis vectors* $e_j \in \mathbb{R}^d$ *for* $j = 1, 2, \dots, d$. *For a triangulation* $\mathcal{T}_h$ *of* Ω *and every* $T \in \mathcal{T}_h$, *there exists an affine diffeomorphism* $\Phi_T : \widehat{T} \to T$, $\Phi_T(\widehat{x}) = B\widehat{x} + b$, *with*

$$\max_{i,j=1,\dots,d} |b_{ij}| \le c h_T, \quad \max_{i,j=1,\dots,d} |b_{ij}^{(-1)}| \le c\rho_T^{-1}$$

for the entries b_{ij} *and* $b_{ij}^{(-1)}$ *of* B *and* B^{-1}, $i, j = 1, 2, \dots, d$. *For* $v \in W^{m,p}(T)$ *and* $\widehat{v} = v \circ \Phi_T \in W^{m,p}(\widehat{T})$, *we have*

$$|v|_{W^{k,p}(T)} \le c\rho_T^{-k} |\det B|^{1/p} |\widehat{v}|_{W^{k,p}(\widehat{T})}, \quad |\widehat{v}|_{W^{k,p}(\widehat{T})} \le c h_T^{-k} |\det B|^{-1/p} |v|_{W^{k,p}(T)};$$

in particular $|v - I_T v|_{W^{k,p}(T)} \le c_I (h_T^m / \rho_T^k) |v|_{W^{m,p}(T)}$.

Proof The proof follows from the transformation formula

$$\int_T |D^\alpha v|^p \,\mathrm{d}x = |\det D\Phi_T|^p \int_{\widehat{T}} |(D^\alpha v) \circ \Phi_T|^p \,\mathrm{d}x$$

and analogous identities for $\widehat{v}$. □

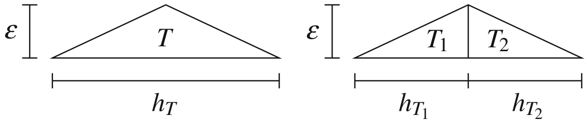


Fig. 3.3 A *triangle* that violates the minimum angle condition if $\varepsilon/h_T \to 0$ (*left*) and *triangles* that satisfy the maximum angle condition even for $\varepsilon/h_{T_\ell} \to 0$, $\ell = 1, 2$ (*right*)

Definition 3.4 A family of (conforming) triangulations $(\mathscr{T}_h)_{h>0}$ is called (*shape*) *regular* if there exists a constant $c > 0$ such that $\sup_{h>0} \sup_{T\in\mathscr{T}_h} h_T/\rho_T \le c$.

The index h in a family of triangulations $(\mathscr{T}_h)_{h>0}$ usually refers to a characteristic or maximal size of the elements in $\mathscr{T}_h$, e.g., it is typically assumed that $\max_{T\in\mathscr{T}_h} h_T \le ch$ for all $h > 0$. Nevertheless, for a sequence of locally refined triangulations, we may have $\max_{T\in\mathscr{T}_h} h_T = \max_{T'\in\mathscr{T}_{h'}} h_{T'}$ for two different triangulations $\mathscr{T}_h$ and $\mathscr{T}_{h'}$. In this case h may refer to an average mesh-size.

Remark 3.1 For shape regularity, the *minimum angle condition*, requiring that the angles of triangles be uniformly bounded from below by a positive number, is sufficient. A weaker *maximum angle condition* is sufficient for a robust interpolation estimate.

Example 3.2 We consider the triangulations $\mathscr{T}_1$ and $\mathscr{T}_2$ displayed in Fig. 3.3 and the function $u(x_1, x_2) = 1 - x_1^2$ for $x = (x_1, x_2) \in \mathbb{R}^2$. For the triangulation $\mathscr{T}_1^\varepsilon = \{T\}$ with

$$T = \operatorname{conv}\{(-1, 0), (1, 0), (0, \varepsilon)\},$$

we have $\mathscr{I}_1 u(x_1, x_2) = x_2/\varepsilon$. For the triangulation $\mathscr{T}_2^\varepsilon = \{T_1, T_2\}$ with

$$T_1 = \operatorname{conv}\{(-1, 0), (0, 0), (0, \varepsilon)\}, \quad T_2 = \operatorname{conv}\{(0, 0), (1, 0), (0, \varepsilon)\},$$

we have $\mathscr{I}_2 u(x_1, x_2) = 1 - |x_1|$.

Definition 3.5 For a triangulation $\mathscr{T}_h$, we let $\mathscr{N}_h$ denote the set of vertices of elements called *nodes* and $\mathscr{S}_h$ to be the set of $(d-1)$-dimensional sides of elements in $\mathscr{T}_h$, i.e., endpoints of intervals, edges of triangles, or faces of tetrahedra if $d = 1, 2, 3$, respectively.

The notation is illustrated in Fig. 3.4.

Definition 3.6 The *P*1*-finite element space* subordinated to a triangulation $\mathscr{T}_h$ is the space

$$\mathscr{S}^1(\mathscr{T}_h) = \{v_h \in C(\overline{\Omega}) : v_h|_T \in P_1(T) \text{ for all } T \in \mathscr{T}_h\}.$$

The subset of functions in $\mathscr{S}^1(\mathscr{T}_h)$, satisfying homogeneous Dirichlet conditions on a subset $\Gamma_{\mathrm{D}} \subset \partial\Omega$, is defined as

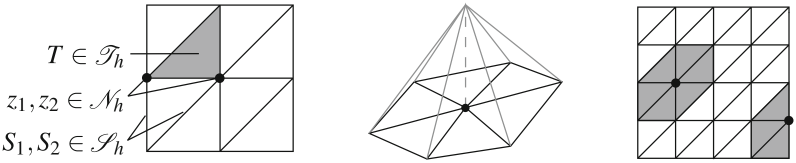


Fig. 3.4 Element $T \in \mathscr{T}_h$, nodes $z_1, z_2 \in \mathscr{N}_h$, and sides $S_1, S_2 \in \mathscr{S}_h$ (*left*), nodal basis functions φ_z (*middle*), and supports ω_z of nodal basis functions φ_z for different nodes $z \in \mathscr{N}_h$ (*right*)

$$\mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h) = \mathscr{S}^1(\mathscr{T}_h) \cap H^1_{\mathrm{D}}(\Omega).$$

If $\Gamma_{\mathrm{D}} = \partial\Omega$, we also write $\mathscr{S}^1_0(\mathscr{T}_h)$ instead of $\mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$. The *nodal basis* of $\mathscr{S}^1(\mathscr{T}_h)$ is the family $(\varphi_z : z \in \mathscr{N}_h)$ with functions $\varphi_z \in \mathscr{S}^1(\mathscr{T}_h)$ satisfying $\varphi_z(y) = \delta_{zy}$ for all $z, y \in \mathscr{N}_h$. The *nodal interpolant* of a function $v \in C(\overline{\Omega})$ is defined by

$$\mathscr{I}_h v = \sum_{z \in \mathscr{N}_h} v(z)\varphi_z.$$

Theorem 3.2 (Nodal interpolation estimates) *For a regular family of triangulations* $(\mathscr{T}_h)_{h>0}$ *such that* $\max_{T\in\mathscr{T}_h} h_T \le ch$ *and* $v \in W^{2,p}(\Omega)$, *we have that* $\mathscr{I}_h v \in \mathscr{S}^1(\mathscr{T}_h)$, *and for every* $1 \le p \le \infty$ *with* $p > d/2$ *if* $d \ge 3$, *we have*

$$h^{-1}\|v - \mathscr{I}_h v\|_{L^p(\Omega)} + \|\nabla(v - \mathscr{I}_h v)\|_{L^p(\Omega)} \le ch\|D^2 v\|_{W^{2,p}(\Omega)}.$$

Moreover, if $v|_{\Gamma_{\mathrm{D}}} = 0$, *then* $\mathscr{I}_h v|_{\Gamma_{\mathrm{D}}} = 0$.

Proof Estimates follow from the stability of interpolation and the transformation estimates if $1 \le p < \infty$. The case $p = \infty$ is treated directly using that functions in $W^{1,\infty}(\Omega)$ are Lipschitz continuous. □

Remark 3.2 For $p = \infty$ we also have $\|v - \mathscr{I}_h v\|_{L^\infty(\Omega)} \le ch\|\nabla v\|_{L^\infty(\Omega)}$.

3.1.3 Projection and Quasi-Interpolation Operators

The nodal interpolation operator $\mathscr{I}_h$ can only be applied to continuous functions and this is often too restrictive in practice. A way to avoid this is to regularize a function by mollification, but this is also often not practical. The difficulties can be circumvented by using projection and quasi-interpolation operators. We assume that Γ_{D} has positive surface measure.

Definition 3.7 The L^2- *and* H^1*-projections* of functions $v \in L^2(\Omega)$ and $w \in H^1_{\mathrm{D}}(\Omega)$ are the uniquely defined functions $P_h v \in \mathscr{S}^1(\mathscr{T}_h)$ and $Q_h w \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ that satisfy

$$\int_\Omega (P_h v - v)\phi_h \,\mathrm{d}x = 0, \qquad \int_\Omega \nabla(Q_h w - w)\cdot\nabla\psi_h \,\mathrm{d}x = 0$$

for all $\phi_h \in \mathscr{S}^1(\mathscr{T}_h)$ and all $\psi_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$, respectively.

Remark 3.3 The operators P_h and Q_h are linear and bounded as operators on $L^2(\Omega)$ and $H^1_{\mathrm{D}}(\Omega)$ with operator norm 1, respectively. They are equivalently characterized by the best-approximation properties

$$\begin{aligned} \|v - P_h v\| &= \min_{\phi\in\mathscr{S}^1(\mathscr{T}_h)} \|v - \phi\|, \\ \|\nabla(w - Q_h w)\| &= \min_{\psi_h\in\mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)} \|\nabla(w - \psi_h)\|. \end{aligned}$$

In the absence of Dirichlet boundary conditions, the H^1-norm can be used instead of the seminorm to define Q_h.

Lemma 3.1 (Projection error) *If $w \in H^2(\Omega) \cap H^1_{\mathrm{D}}(\Omega)$, then we have*

$$\|w - P_h w\| \le ch^2\|D^2 w\|, \quad \|\nabla(w - Q_h w)\| \le ch\|D^2 w\|.$$

Proof The estimates follow from the best-approximation properties and estimates for nodal interpolation. □

Remark 3.4 We will show below that under certain conditions on Ω, the Aubin–Nitsche lemma implies $\|w - Q_h w\| \le ch^2\|D^2 w\|$.

The operators P_h and Q_h can be applied to a large class of possibly discontinuous functions and their orthogonality property is important in many estimates. A disadvantage is the global character of their construction. An intermediate solution between interpolation and projection is provided by quasiinterpolants.

Definition 3.8 The *Clément interpolant* $\mathscr{J}_h v \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ of a function $v \in L^1(\Omega)$ is defined by $\mathscr{J}_h v = \sum_{z\in\mathscr{N}_h} v_z\varphi_z$, where

$$v_z = \begin{cases} |\omega_z|^{-1}\int_{\omega_z} v\,\mathrm{d}x & \text{if} \quad z \in \mathscr{N}_h\setminus\Gamma_{\mathrm{D}}, \\ 0 & \text{if} \quad z \in \mathscr{N}_h \cap \Gamma_{\mathrm{D}}, \end{cases}$$

with the *node patch* $\omega_z = \operatorname{supp}\varphi_z$ for every $z \in \mathscr{N}_h$ with diameter $h_z = \operatorname{diam}(\omega_z)$.

Remark 3.5 The coefficients $(v_z)_{z\in\mathscr{N}_h}$ in the definition of $\mathscr{J}_h v$ are equivalently defined by local projections onto constants, i.e., for every $z \in \mathscr{N}_h\setminus\Gamma_{\mathrm{D}}$ we have that v_z is the unique minimum of the mapping $c \mapsto \|v - c\|^2_{L^2(\omega_z)}$.

Some local estimates are required to analyze the approximation properties of $\mathscr{J}_h$.

Lemma 3.2 (Local Poincaré inequality) *There exists $c > 0$ such that for all $z \in \mathcal{N}_h$ and $v \in H^1_{\mathrm{D}}(\Omega)$, we have*

$$\|v - v_z\|_{L^2(\omega_z)} \le c h_z \|\nabla v\|_{L^2(\omega_z)}.$$

The constant c depends on the shapes of the sets $(\omega_z : z \in \mathcal{N}_h)$.

Proof Assume first that $h_z = 1$. If $z \in \mathcal{N}_h \setminus \Gamma_{\mathrm{D}}$, then we have $\int_{\omega_z}(v - v_z)\,\mathrm{d}x = 0$ and the Poincaré inequality implies that $\|v - v_z\|_{L^2(\omega_z)} \le c\|\nabla v\|_{L^2(\omega_z)}$. If $z \in \mathcal{N}_h \cap \Gamma_{\mathrm{D}}$ then $\omega_z \cap \Gamma_{\mathrm{D}}$ has positive surface measure and the estimate follows from Friedrichs' inequality. A transformation argument shows the dependence on h_z. □

Lemma 3.3 (Trace inequality) *Let $T \in \mathcal{T}_h$ and $S \in \mathcal{S}_h$ such that $S \subset \partial T$. There exists $c > 0$ such that for all $v \in H^1(T)$, we have*

$$\|v\|_{L^2(S)} \le c\big(h_S^{-1/2}\|v\|_{L^2(T)} + h_S^{1/2}\|\nabla v\|_{L^2(T)}\big).$$

Proof The proof uses the density of smooth functions and a one-dimensional integration-by-parts formula to express function values on S by integrals over line segments in T. □

Remark 3.6 For a regular family of triangulations, there exists an h-independent constant $c > 0$ such that $c^{-1}h_T \le h_z \le c h_T$ if $z \in \mathcal{N}_h$ and $T \in \mathcal{T}_h$ with $z \in T$. If the triangulations are nested, i.e., obtained by successive refinement, then only a finite number of shapes of patches can occur.

Theorem 3.3 (Clément interpolation) *There exists $c > 0$ such that for all $v \in H^1_{\mathrm{D}}(\Omega)$, we have*

$$\|\nabla \mathcal{J}_h v\| + \|h_{\mathcal{T}}^{-1}(v - \mathcal{J}_h v)\| + \|h_{\mathcal{S}}^{-1/2}(v - \mathcal{J}_h v)\|_{L^2(\cup\mathcal{S}_h)} \le c\|\nabla v\|,$$

where $h_{\mathcal{T}} \in L^\infty(\Omega)$ is defined by $h_{\mathcal{T}}|_T = h_T$ and $h_{\mathcal{S}} \in L^\infty(\cup\mathcal{S}_h)$ by $h_{\mathcal{S}}|_S =$ $\mathrm{diam}(S)$ for every $S \in \mathcal{S}_h$.

Proof The nodal basis functions form a partition of unity, i.e., $\sum_{z\in\mathcal{N}_h}\varphi_z = 1$ almost everywhere in Ω, with finite overlap. Moreover, we have $\|\varphi_z\|_{L^\infty(\omega_z)} = 1$ and $\|\nabla\varphi_z\|_{L^\infty(\omega_z)} \le c h_z^{-1}$ for every $z \in \mathcal{N}_h$. Using $\sum_{z\in\mathcal{N}_h}\nabla\varphi_z = 0$ and the local Poincaré inequality, we have

$$\begin{aligned}
\|\nabla \mathcal{J}_h v\|^2 &= \int_\Omega \sum_{z\in\mathcal{N}_h}(v_z - v)\nabla\varphi_z \cdot \nabla \mathcal{J}_h v\,\mathrm{d}x \\
&\le \sum_{z\in\mathcal{N}_h}\|v_z - v\|_{L^2(\omega_z)}\|\nabla\varphi_z\|_{L^\infty(\omega_z)}\|\nabla \mathcal{J}_h v\|_{L^2(\omega_z)} \\
&\le c\sum_{z\in\mathcal{N}_h}\|\nabla v\|_{L^2(\omega_z)}\|\nabla \mathcal{J}_h v\|_{L^2(\omega_z)}
\end{aligned}$$

$$\leq c\Big(\sum_{z\in\mathscr{N}_h}\|\nabla v\|^2_{L^2(\omega_z)}\Big)^{1/2}\Big(\sum_{z\in\mathscr{N}_h}\|\nabla\mathscr{J}_h v\|^2_{L^2(\omega_z)}\Big)^{1/2}$$
$$\leq c\|\nabla v\|\|\nabla\mathscr{J}_h v\|$$

which is the first estimate. To prove the second estimate, we let $\psi\in L^2(\Omega)$ and note that we have

$$\begin{aligned}\int_\Omega (v-\mathscr{J}_h v)\psi\,\mathrm{d}x &= \sum_{z\in\mathscr{N}_h}\int_{\omega_z}\varphi_z(v-v_z)\psi\,\mathrm{d}x\\ &\leq \sum_{z\in\mathscr{N}_h}\|\varphi_z\|_{L^2(\omega_z)}\|v-v_z\|_{L^2(\omega_z)}\|\psi\|_{L^2(\omega_z)}\\ &\leq c\Big(\sum_{z\in\mathscr{N}_h}\|\nabla v\|^2_{L^2(\omega_z)}\Big)^{1/2}\Big(\sum_{z\in\mathscr{N}_h}h_z\|\psi\|^2_{L^2(\omega_z)}\Big)^{1/2}\\ &\leq c\|\nabla v\|\|h_{\mathscr{T}}\psi\|.\end{aligned}$$

The choice of $\psi = h_{\mathscr{T}}^{-2}(v-\mathscr{J}_h v)$ implies the second estimate. With the trace inequality we verify that for every $S\in\mathscr{S}_h$ with neighboring element $T_S\in\mathscr{T}_h$, we have

$$c^{-1}h_S^{-1}\|v-\mathscr{J}_h v\|^2_{L^2(S)}\leq h_S^{-1}\|v-\mathscr{J}_h v\|^2_{L^2(T_S)}+h_S\|\nabla(v-\mathscr{J}_h v)\|^2_{L^2(T_S)}.$$

A summation over $S\in\mathscr{S}_h$ combined with the first two estimates imply the estimate. □

Remarks 3.7 (i) The Clément interpolant $\mathscr{J}_h$ is not a projection operator, i.e., in general we have $\mathscr{J}_h v_h \neq v_h$ for $v_h\in\mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$. Certain modifications of the operator guarantee this property.
(ii) The constant in the theorem remains bounded for a regular family of nested triangulations.
(iii) The Bramble–Hilbert lemma implies the local approximation estimate

$$\|v-\mathscr{J}_j v\|_{L^2(T)}+h_T^{-1}\|\nabla(v-\mathscr{J}_h v)\|_{L^2(T)}\leq ch_T^2\|D^2 v\|_{L^2(\omega_T)}$$

for $v\in H^2(\omega_T)$ with $\omega_T=\cup_{z\in\mathscr{N}_h\cap\mathscr{T}_h}\omega_z$.
(iv) The estimates remain valid for $\Gamma_{\mathrm{D}}=\emptyset$ and exponents $p\in(1,\infty)$.

3.1.4 Other Estimates

We collect some useful estimates for functions in $\mathscr{S}^1(\mathscr{T}_h)$.

Lemma 3.4 (Norm equivalence) *For every* $1 \le p < \infty$ *there exists* $c > 0$ *such that for all* $v_h \in \mathscr{S}^1(\mathscr{T}_h)$, *we have*

$$c^{-1}\|v_h\|_{L^p(\Omega)} \le \Big(\sum_{z\in\mathscr{N}_h} h_z^d |v_h(z)|^p\Big)^{1/p} \le c\|v_h\|_{L^p(\Omega)}.$$

Moreover, we have $\|v_h\|_{L^\infty(\Omega)} = \max_{z\in\mathscr{N}_h} |v_h(z)|$ *for every* $v_h \in \mathscr{S}^1(\mathscr{T}_h)$.

Proof For every $T \in \mathscr{T}_h$ the expressions $\|v_h\|_{L^p(T)}$ and $\big(h_T^d \sum_{z\in\mathscr{N}_h\cap T} |v_h(z)|^p\big)^{1/p}$ are norms on the finite-dimensional space $\mathscr{S}^1(\mathscr{T}_h)|_T$. Hence they are equivalent and a transformation argument shows that the constant is independent of h_T and h_z. The asserted estimate follows from a summation over $T \in \mathscr{T}_h$. □

Definition 3.9 A family of triangulations $(\mathscr{T}_h)_{h>0}$ is called *quasiuniform* if there exists $c > 0$ such that $c^{-1}h \le h_T \le ch$ for all $h > 0$ and all $T \in \mathscr{T}_h$.

Lemma 3.5 (Inverse estimates) *For* $v_h \in \mathscr{S}^1(\mathscr{T}_h)$ *and* $1 \le r, p \le \infty$ *we have*

$$\|\nabla v_h\|_{L^p(T)} \le ch_T^{-1}\|v_h\|_{L^p(T)}$$

and

$$\|v_h\|_{L^p(T)} \le ch_T^{d(r-p)/(pr)}\|v_h\|_{L^r(T)}.$$

If the family $(\mathscr{T}_h)_{h>0}$ *is quasiuniform, then we have*

$$\|\nabla v_h\|_{L^p(\Omega)} \le ch^{-1}\|v_h\|_{L^p(\Omega)}$$

and

$$\|v_h\|_{L^p(\Omega)} \le ch^{\min\{0,d(r-p)/(pr)\}}\|v_h\|_{L^r(\Omega)}$$

Proof To prove the first estimate, consider the space $\mathscr{S}^1(\mathscr{T}_h)|_T/\mathbb{R}$, i.e., functions $v_h \in \mathscr{S}^1(\mathscr{T}_h)$ with $\int_T v_h \,\mathrm{d}x = 0$. The expressions $\|\nabla v_h\|_{L^p(T)}$ and $h_T^{-1}\|v_h\|_{L^p(T)}$ are equivalent norms on the finite-dimensional space $\mathscr{S}^1(\mathscr{T}_h)|_T/\mathbb{R}$. Using the estimate $\|v_h - \overline{v}_h\|_{L^p(T)} \le \|v_h\|_{L^p(T)}$ for $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ and $\overline{v}_h = |T|^{-1}\int_T v_h \,\mathrm{d}x$, a transformation argument proves the first estimate. A similar argument proves the second estimate. The third estimate follows from a summation of the first estimate over $T \in \mathscr{T}_h$ and $h_T^{-1} \le ch^{-1}$ due to the assumed quasiuniformity of $\mathscr{T}_h$. To prove the last estimate we first note that it follows directly from Hölder's inequality if $p \ge r$. Otherwise, we use $\big(\sum_{j=1}^L |x_j|^r\big)^{1/r} \le \big(\sum_{j=1}^L |x_j|^p\big)^{1/p}$ for every $L \in \mathbb{N}$ and $x \in \mathbb{R}^L$ and deduce that

$$\Big(\sum_{T\in\mathscr{T}_h} \|v_h\|_{L^r(T)}^p\Big)^{1/p} \le \Big(\sum_{T\in\mathscr{T}_h} \|v_h\|_{L^r(T)}^r\Big)^{1/r} = \|v_h\|_{L^r(\Omega)}.$$

With the corresponding elementwise estimates, this implies the global estimate for quasiuniform triangulations. □

Remark 3.8 For quasiuniform triangulations we also have the inverse estimate

$$\|v_h\|_{L^\infty(\Omega)} \le ch^{1-d/2} \log h^{-1} \|v_h\|_{W^{1,2}(\Omega)}.$$

A proof follows from the Sobolev estimate $\|v\|_{L^p(\Omega)} \le cp\|v\|_{W^{1,q}(\Omega)}$ for $1 \le q < d$ and $p = dq/(d-q)$, the choice of $q = d - |\log h|^{-1}$, and the inverse estimates of the lemma.

The union of a family of finite element spaces $\left(\mathscr{S}^1(\mathscr{T}_h)\right)_{h>0}$ is dense in $W^{1,p}(\Omega)$ for $1 \le p < \infty$.

Lemma 3.6 (Density) *For* $1 \le p < \infty$ *and* $v \in W^{1,p}(\Omega)$ *there exists a sequence* $(v_h)_{h>0} \subset W^{1,p}(\Omega)$ *with* $v_h \in \mathscr{S}^1(\mathscr{T}_h)$ *for every* $h > 0$ *such that* $v_h \to v$ *in* $W^{1,p}(\Omega)$ *as* $h \to 0$.

Proof Assume that $1 \le p < \infty$ with $p > d/2$ if $d \ge 3$. The set $C^\infty(\overline{\Omega}) \cap W^{1,p}(\Omega)$ is dense in $W^{1,p}(\Omega)$ and for $\varepsilon > 0$ we may choose $v_\varepsilon \in C^\infty(\overline{\Omega}) \cap W^{1,p}(\Omega)$ such that $\|v - v_\varepsilon\|_{W^{1,p}(\Omega)} \le \varepsilon/2$ and $\|D^2 v_\varepsilon\|_{L^p(\Omega)} \le c\varepsilon^{-1}\|\nabla v\|_{L^p(\Omega)}$. Setting $v_h = \mathscr{I}_h v_\varepsilon$, we have $\|v_h - v_\varepsilon\|_{W^{1,p}(\Omega)} \le h\|D^2 v_\varepsilon\|_{L^p(\Omega)}$ and for h sufficiently small, it follows that $\|v_h - v\|_{W^{1,p}(\Omega)} \le \varepsilon$. If $p \le d/2$ we may use that $\|D^2 v_\varepsilon\| \le c\varepsilon^{-1-d/p'}\|v\|_{W^{1,p}(\Omega)}$ to verify the statement. □

With the density of finite element functions it follows that projections satisfy a super-approximation property.

Corollary 3.2 (Super-approximation) *For every* $v \in H^1(\Omega)$ *we have* $\|v - P_h v\| = o(h)$ *as* $h \to 0$*, i.e., for every* $\varepsilon > 0$ *there exists* $h_0 > 0$ *such that* $\|v - P_h v\| \le \varepsilon h$ *for all* $0 < h \le h_0$.

Proof Let $\varepsilon > 0$. The difference $v - P_h v$ is orthogonal to the subspace $\mathscr{S}^1(\mathscr{T}_h)$ in $L^2(\Omega)$ and therefore we have

$$\|v - P_h v\|^2 = \int_\Omega (v - P_h v)(v - P_h v - w_h)\,\mathrm{d}x$$

for all $w_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$. Because of the previous lemma there exists $h_0 > 0$ and $v_h \in \mathscr{S}^1(\mathscr{T}_h)$ such that $\|\nabla(v - v_h)\| \le \varepsilon$ for all $0 < h \le h_0$. With the choice of the function $w_h = v_h - P_h v + \mathscr{J}_h(v - v_h)$ we have

$$\|v - P_h v\|^2 \le \|v - P_h v\| \|(v - v_h) - \mathscr{J}_h(v - v_h)\|.$$

With the estimates for the Clément interpolant (generalized to the case $\Gamma_{\mathrm{D}} = \emptyset$) we deduce

$$\|(v - v_h) - \mathscr{J}_h(v - v_h)\| \leq ch\|\nabla(v - v_h)\|$$

and the combination of the estimates implies the statement. □

For a polynomial function, a Poincaré inequality holds if the function vanishes at a single point.

Lemma 3.7 (Discrete Poincaré inequality) *Let $T \in \mathscr{T}_h$, $z \in T \cap \mathscr{N}_h$, $1 \leq p \leq \infty$, and $k \in \mathbb{N}$. There exists a constant $c_T > 0$ that is independent of the diameter of T such that for all $v_h \in P_k(T)$ with $v_h(z) = 0$, we have*

$$\|v_h\|_{L^p(T)} \leq ch_T\|\nabla v_h\|_{L^p(T)}.$$

Proof If $k = 1$, the proof follows from the fact that for $x \in T$, we have

$$v_h(x) = v_h(z) + \nabla v_h|_T \cdot (x - z).$$

If $k \geq 1$, we argue by contradiction and let $(w_h^j)_{j\in\mathbb{N}}$ be a sequence in $P_k(T)$ such that $w_h^j(z) = 0$ for all $j \in \mathbb{N}$ and $1 = \|w_h^j\|_{L^p(T)} > jh_T\|\nabla w_h^j\|_{L^p(T)}$. The bounded sequence $(w_h^j)_{j\in\mathbb{N}}$ has a convergent subsequence with limit $w_h \in P_k(T)$ satisfying $w_h(z) = 0$. The triangle inequality and an inverse estimate imply that $\|\nabla w_h\|_{L^p(T)} = 0$, i.e., that w_h is constant with value 0. This contradicts $\|w_h\|_{L^p(T)} = \lim_{j\to\infty}\|w_h^j\|_{L^p(T)} = 1$. Hence there exists a constant $c > 0$, so that the asserted estimate holds. A scaling argument proves that c is independent of h_T. □

3.2 Approximation of the Poisson Problem

Given a bounded, polyhedral Lipschitz domain $\Omega \subset \mathbb{R}^d$, a closed subset $\Gamma_{\mathrm{D}} \subset \partial\Omega$ with positive surface measure, $u_{\mathrm{D}} \in C(\Gamma_{\mathrm{D}})$ with $u_{\mathrm{D}} = \widetilde{u}_{\mathrm{D}}|_{\Gamma_{\mathrm{D}}}$ for some $\widetilde{u}_{\mathrm{D}} \in H^1(\Omega)$, $g \in L^2(\Gamma_{\mathrm{N}})$, and $f \in L^2(\Omega)$, we consider the boundary value problem

$$-\Delta u = f \text{ in } \Omega, \quad u|_{\Gamma_{\mathrm{D}}} = u_{\mathrm{D}}, \quad \partial_n u|_{\Gamma_{\mathrm{N}}} = g.$$

By decomposing $u = \widetilde{u} + \widetilde{u}_{\mathrm{D}}$ with $\widetilde{u} \in H^1_{\mathrm{D}}(\Omega)$ and replacing f and g by $f + \Delta\widetilde{u}_{\mathrm{D}}$ and $g - \partial_n\widetilde{u}_{\mathrm{D}}$, respectively, provided $\widetilde{u}_{\mathrm{D}} \in H^2(\Omega)$ we may and will assume that $u_{\mathrm{D}} = 0$ unless stated otherwise. Within this setting we review standard concepts for the numerical analysis of finite element methods for the elliptic model problem, and refer the reader to [2, 3, 11] for further details.

3.2.1 Variational Formulation

The boundary value problem is the strong form of the Euler–Lagrange equations of the minimization problem defined by the functional

$$I(u) = \frac{1}{2}\int\limits_{\Omega} |\nabla u|^2 \,\mathrm{d}x - \int\limits_{\Omega} fu \,\mathrm{d}x - \int\limits_{\Gamma_\mathrm{N}} gu \,\mathrm{d}s$$

for $u \in H^1_\mathrm{D}(\Omega)$, and the direct method in the calculus of variations implies the existence of a unique solution $u \in H^1_\mathrm{D}(\Omega)$. The weak form of the Euler–Lagrange equations states that a minimizer $u \in H^1_\mathrm{D}(\Omega)$ satisfies

$$\int\limits_{\Omega} \nabla u \cdot \nabla v \,\mathrm{d}x = \int\limits_{\Omega} fv \,\mathrm{d}x + \int\limits_{\Gamma_\mathrm{D}} gv \,\mathrm{d}s$$

for all $v \in H^1_\mathrm{D}(\Omega)$. Equivalently, the *Lax-Milgram lemma* shows the existence of a unique solution of the weak form of the Euler-Lagrange equations. For this, it suffices to realize, with the help of Poincaré and Hölder inequalities, that the bilinear form defined for $v, w \in H^1_\mathrm{D}(\Omega)$ by

$$a(v, w) = \int\limits_{\Omega} \nabla v \cdot \nabla w \,\mathrm{d}x$$

is bounded and coercive on $H^1_\mathrm{D}(\Omega) \times H^1_\mathrm{D}(\Omega)$ and that the right-hand side of the weak formulation defines a bounded linear functional on $H^1_\mathrm{D}(\Omega)$.

Theorem 3.4 (Existence and stability) *There exists a unique minimizer $u \in H^1_\mathrm{D}(\Omega)$ of the functional I which solves the weak form of the Euler–Lagrange equations and satisfies*

$$\|u\|_{H^1(\Omega)} \le c(\|f\| + \|g\|_{L^2(\Gamma_\mathrm{N})}).$$

The theorem implies that the solution operator is bounded as a mapping $L^2(\Omega) \times L^2(\Gamma_\mathrm{N}) \to H^1(\Omega)$. In certain situations the solution operator attains its values in $H^1_\mathrm{D}(\Omega) \cap H^2(\Omega)$.

Definition 3.10 The Poisson problem with homogeneous Dirichlet boundary conditions is called *H^2-regular* if there exists a constant $c > 0$ such that

$$\|u\|_{H^2(\Omega)} \le c(\|f\| + \|g\|_{L^2(\Gamma_\mathrm{N})}).$$

Example 3.3 If $\Omega \subset \mathbb{R}^2$ is convex and $\Gamma_\mathrm{D} = \partial\Omega$, then the Poisson problem is H^2-regular.

3.2.2 Error Estimates

For a shape-regular but not necessarily quasiuniform family of conforming triangulations, the *Galerkin approximation* of the Poisson problem is for every $h > 0$ defined as the minimizer of the energy functional I restricted to $\mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$, or equivalently as the unique function $u_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ that satisfies

$$\int_{\Omega} \nabla u_h \cdot \nabla v_h \,\mathrm{d}x = \int_{\Omega} f v_h \,\mathrm{d}x + \int_{\Gamma_{\mathrm{N}}} g v_h \,\mathrm{d}s$$

for all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$. Existence and uniqueness of u_h are direct consequences of the Lax–Milgram lemma. An important property of the Galerkin approximation $u_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ is that the approximation error $u - u_h$ satisfies the *Galerkin orthogonality*

$$\int_{\Omega} \nabla(u - u_h) \cdot \nabla v_h \,\mathrm{d}x = 0$$

for all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$. The interpretation of this identity is that $u_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ is the H^1-projection of the exact solution $u \in H^1_{\mathrm{D}}(\Omega)$ onto the subspace $\mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$. In particular, it satisfies a (quasi-) *best-approximation* property or, more generally, the conditions of *Céa's lemma* are satisfied, i.e., we have

$$\|\nabla(u - u_h)\| \le \inf_{v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)} \|\nabla(u - v_h)\|.$$

The density of finite element spaces in $H^1_{\mathrm{D}}(\Omega)$ implies convergence $u_h \to u$ in $H^1_{\mathrm{D}}(\Omega)$ as $h \to 0$ and if $u \in H^2(\Omega)$, we obtain a convergence rate.

Corollary 3.3 (Approximation error) *If* $u \in H^2(\Omega) \cap H^1_{\mathrm{D}}(\Omega)$, *then we have*

$$\|\nabla(u - u_h)\| \le ch\|D^2 u\|.$$

Proof The error estimate follows from the best-approximation property and the nodal interpolation estimates. □

Remarks 3.9 (i) The error estimate is special due to the fact that the approximation u_h is the H^1-projection of the exact solution, i.e., $u_h = Q_h u$. A more general concept is based on a consistency property of the discretization and a stability estimate for the numerical method.
(ii) For $w_h \in \mathscr{S}^1_0(\mathscr{T}_h)$ the discrete Laplace operator $-\Delta_h w_h \in \mathscr{S}^1_0(\mathscr{T}_h)$ is the uniquely defined function that satisfies

$$(-\Delta_h w_h, v_h) = (\nabla v_h, \nabla w_h)$$

for all $w_h \in \mathscr{S}_0^1(\mathscr{T}_h)$. In particular, if $\Gamma_{\rm D} = \partial\Omega$, the Galerkin approximation $u_h \in \mathscr{S}_0^1(\mathscr{T}_h)$ of the Poisson problem satisfies $-\Delta_h u_h = P_{h,0} f$, where $P_{h,0} f$ is the L^2-projection of f onto $\mathscr{S}_0^1(\mathscr{T}_h)$.

For the proof of optimal error estimates in $L^2(\Omega)$, a stronger assumption than $u \in H^2(\Omega)$ is required, namely that the problem be H^2-regular. In this case, the unique weak solution $z \in H_{\rm D}^1(\Omega)$ of the Poisson problem

$$-\Delta z = e \text{ in } \Omega, \quad z|_{\Gamma_{\rm D}} = 0, \quad \partial_n z = 0 \text{ on } \Gamma_{\rm N}$$

with $e = u - u_h$ is a strong solution and satisfies $\|D^2 z\| \le c\|e\|$. Green's formula and Galerkin orthogonality yield that, for every $z_h \in \mathscr{S}_{\rm D}^1(\mathscr{T}_h)$, we have

$$\int_\Omega e^2 \,{\rm d}x = \int_\Omega e(-\Delta z)\,{\rm d}x = \int_\Omega \nabla e \cdot \nabla z \,{\rm d}x = \int_\Omega \nabla e \cdot \nabla(z - z_h)\,{\rm d}x.$$

With Hölder's inequality, the assumed bound for $\|D^2 z\|$, and the choice $z_h = \mathscr{I}_h z$ we find that

$$\begin{aligned} \|e\|^2 &\le \|\nabla e\| \|\nabla(z - z_h)\| \\ &\le ch\|\nabla e\| \|D^2 z\| \le ch\|\nabla e\| \|e\|. \end{aligned}$$

Incorporating the estimate $\|\nabla e\| \le ch\|D^2 u\|$ proves the following result.

Theorem 3.5 (Aubin–Nitsche lemma) *If the Poisson problem is H^2-regular, then*

$$\|u - u_h\| \le ch^2 \|D^2 u\|.$$

Remarks 3.10 (i) The H^1-error estimate can be written in the form $\|\nabla(u - u_h)\| \le c\|h_{\mathscr{T}} D^2 u\|$ and motivates the use of a small local mesh-size where $D^2 u$ is large. Such a localization is only partially possible for the L^2-error estimate.
(ii) By interpolating Green's function associated to the Poisson problem on $\Omega \subset \mathbb{R}^2$ with $\Gamma_{\rm D} = \partial\Omega$, one can show that if the Poisson problem is H^2-regular, if $\mathscr{T}_h$ is quasiuniform, and if $u \in C^2(\overline{\Omega})$, then we have

$$\|u - u_h\|_{L^\infty(\Omega)} \le ch^2(1 + |\log h|)\|D^2 u\|_{L^\infty(\Omega)}.$$

We close the discussion of approximation errors with an *a posteriori error estimate* that bounds the approximation error in $H^1(\Omega)$ by computable quantities.

Definition 3.11 Given $u_h \in \mathscr{S}^1(\mathscr{T}_h)$ and an interior side $S \in \mathscr{S}_h$, i.e., $S = T_1 \cap T_2$ for distinct $T_1, T_2 \in \mathscr{T}_h$, the *jump of* ∇u_h *across* S *in the normal direction to* S is defined as

$$[\![\nabla u_h \cdot n_S]\!] = \nabla u_h|_{T_1} \cdot n_{T_1,S} + \nabla u_h|_{T_2} \cdot n_{T_2,S},$$

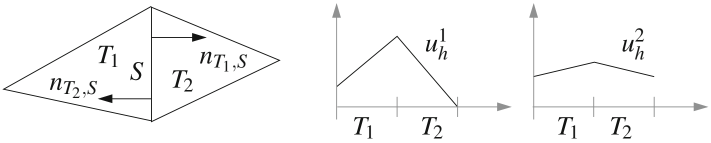


Fig. 3.5 Interior edge $S = T_1 \cap T_2$ and outer unit normals $n_{T_1,S}$ and $n_{T_2,S}$ on S (*left*); large and small jumps of the gradients of functions u_h^1 and u_h^2 (*right*)

where $n_{T_\ell,S}$ is the outer unit normal to T_ℓ on S for $\ell = 1, 2$, cf. Fig. 3.5.

Theorem 3.6 (A posteriori error estimate) *We have*

$$\begin{aligned}(1/c)\|\nabla(u-u_h)\| \le &\Big(\sum_{S\in\mathscr{S}_h\cap\Omega} h_S\|[\![\nabla u_h\cdot n_S]\!]\|^2_{L^2(S)}\Big)^{1/2}\\ &+\Big(\sum_{T\in\mathscr{T}_h} h_T^2\|f+\Delta u_h|_T\|^2_{L^2(T)}\Big)^{1/2}\\ &+\Big(\sum_{S\in\mathscr{S}_h\cap\overline{\Gamma}_{\mathrm{N}}} h_S\|g-\partial_n u_h\|^2_{L^2(S)}\Big)^{1/2}.\end{aligned}$$

Proof We abbreviate $e = u - u_h \in H^1_{\mathrm{D}}(\Omega)$ and note that by Galerkin orthogonality and the properties of the weak solution u, we have

$$\|\nabla e\|^2 = \int_\Omega f(e-\mathscr{J}_h e)\,\mathrm{d}x + \int_{\Gamma_{\mathrm{N}}} g(e-\mathscr{J}_h e)\,\mathrm{d}s - \int_\Omega \nabla u_h\cdot\nabla(e-\mathscr{J}_h e)\,\mathrm{d}x.$$

An elementwise application of Green's formula and a rearrangement of integrals over sides of elements imply that for every $v \in H^1_{\mathrm{D}}(\Omega)$, we have

$$\begin{aligned}-\int_\Omega \nabla u_h\cdot\nabla v\,\mathrm{d}x &= \sum_{T\in\mathscr{T}_h}\Big(\int_T(\Delta u_h|_T)v\,\mathrm{d}x - \int_{\partial T}(\nabla u_h\cdot n_T)v\,\mathrm{d}s\Big)\\ &= \sum_{T\in\mathscr{T}_h}\int_T(\Delta u_h|_T)v\,\mathrm{d}x - \sum_{S\in\mathscr{S}_h\cap\Omega}\int_S[\![\nabla u_h\cdot n_S]\!]v\,\mathrm{d}s\\ &\quad-\int_{\Gamma_{\mathrm{N}}}(\nabla u_h\cdot n)v\,\mathrm{d}s.\end{aligned}$$

With Hölder and Cauchy–Schwarz inequalities, combining previous estimates and the choice of $v = e - \mathscr{J}_h e$ lead to

$$\begin{aligned}\|\nabla e\|^2 \le &\Big(\sum_{T\in\mathscr{T}_h} h_T^2\|f+\Delta u_h|_T\|^2_{L^2(T)}\Big)^{1/2}\Big(\sum_{T\in\mathscr{T}_h} h_T^{-2}\|e-\mathscr{J}_h e\|^2_{L^2(T)}\Big)^{1/2}\\ &+\Big(\sum_{S\in\mathscr{S}_h\cap\overline{\Gamma}_{\mathrm{N}}} h_S\|g-\partial_n u_h\|^2_{L^2(S)}\Big)^{1/2}\Big(\sum_{S\in\mathscr{S}_h\cap\overline{\Gamma}_{\mathrm{N}}} h_S^{-1}\|e-\mathscr{J}_h e\|^2_{L^2(S)}\Big)^{1/2}\\ &+\Big(\sum_{S\in\mathscr{S}_h\cap\Omega} h_S\|[\![\nabla u_h\cdot n_S]\!]\|^2_{L^2(S)}\Big)^{1/2}\Big(\sum_{S\in\mathscr{S}_h\cap\Omega} h_S^{-1}\|e-\mathscr{J}_h e\|^2_{L^2(S)}\Big)^{1/2}.\end{aligned}$$

The approximation properties of the Clément interpolant imply the assertion. □

Remarks 3.11 (i) Note that since $u_h|_T$ is affine, we have $\Delta u_h|_T = 0$ for every $T \in \mathscr{T}_h$. The expression $f + \Delta u_h|_T$ has the interpretation of a residual.
(ii) A converse estimate can be proved up to higher-order terms. This is often called *efficiency* while the estimate of the theorem is referred to as a *reliability* estimate. Note that the estimate holds without regularity requirements on u.
(iii) The upper bound for the error is localizable and leads to strategies for local refinement with quasioptimal convergence rates even if $u \notin H^2(\Omega)$.
(iv) The first term on the right-hand side of the estimate of Theorem 3.6 measures the distance of ∇u_h to the space $H(\mathrm{div};\Omega)$, i.e., the space of square-integrable vector fields that have a weak divergence that is also square-integrable.

3.2.3 Discrete Maximum Principle

The unique minimizer $u \in H^1(\Omega)$ of the Dirichlet energy

$$I(u) = \frac{1}{2}\int_\Omega |\nabla u|^2\,\mathrm{d}x,$$

subject to $u|_{\Gamma_{\mathrm{D}}} = u_{\mathrm{D}}$, satisfies the *maximum principle*

$$\max_{x\in\Omega} u(x) \le \max_{x\in\Gamma_{\mathrm{D}}} u_{\mathrm{D}}(x).$$

A variational proof of this estimate uses the fact that for every $c \in \mathbb{R}$, the truncated function $T_c u(x) = \min\{c, u(x)\}$ belongs to $H^1(\Omega)$ with $\|\nabla T_c u\| \le \|\nabla u\|$. For $m = \max_{x\in\partial\Omega} u_{\mathrm{D}}(x)$ we have $T_m u \le m$, $T_m u|_{\Gamma_{\mathrm{D}}} = u_{\mathrm{D}}$, and

$$I(T_m u) \le I(u).$$

Since u is minimal, we conclude that $T_m u = u$ and $u \le m$ in Ω. This argument cannot be transferred directly to finite element approximations since the truncation $T_c u_h$ of $u_h \in \mathscr{S}^1(\mathscr{T}_h)$ is in general not contained in $\mathscr{S}^1(\mathscr{T}_h)$, cf. Fig. 3.6. Additional conditions have to be imposed to guarantee a discrete version of the maximum

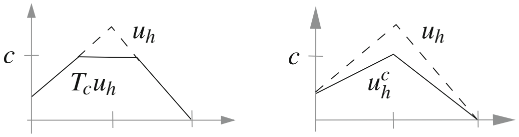


Fig. 3.6 Truncated finite element function $T_c u_h$ that does not belong to the finite element space (*left*) and the function u_h^c that is obtained by a truncation of the nodal values which belongs to the finite element space (*right*); note that $u_h^c = \mathscr{I}_h(T_c u_h)$

principle. The following result provides a discrete version of the estimate $\|\nabla(F \circ v)\| \le \|DF\|_{L^\infty(\mathbb{R})}\|\nabla v\|$ for $v \in H^1(\Omega)$ and a Lipschitz continuous function $F \in W^{1,\infty}(\mathbb{R})$.

Proposition 3.2 (Lipschitz stability) *Assume that the triangulation $\mathscr{T}_h$ of Ω is such that the* stiffness matrix *satisfies*

$$A_{zy} = \int_\Omega \nabla\varphi_z \cdot \nabla\varphi_y \,\mathrm{d}x \le 0$$

for all distinct $z, y \in \mathscr{N}_h$. Then for every $v_h \in \mathscr{S}^1(\mathscr{T}_h)^m$, $F \in W^{1,\infty}(\mathbb{R}^m;\mathbb{R}^\ell)$, and $v_h^F = \mathscr{I}_h(F \circ v) \in \mathscr{S}^1(\mathscr{T}_h)^\ell$, i.e.,

$$v_h^F = \sum_{z\in\mathscr{N}_h} F(v_h(z))\varphi_z,$$

we have

$$\|\nabla v_h^F\| \le \|DF\|_{L^\infty(\mathbb{R}^m)}\|\nabla v_h\|.$$

Proof We set $A_{zy} = \int_\Omega \nabla\varphi_z \cdot \nabla\varphi_y \,\mathrm{d}x$ for all $z, y \in \mathscr{N}_h$ and note that $A_{zy} = A_{yz}$ and for every $y \in \mathscr{N}_h$, we have

$$\sum_{z\in\mathscr{N}_h} A_{zy} = \sum_{z\in\mathscr{N}_h}\int_\Omega \nabla\varphi_z \cdot \nabla\varphi_y \,\mathrm{d}x = 0$$

due to the fact that $\sum_{y\in\mathscr{N}_h}\varphi_y = 1$ in Ω. For $w_h = \sum_{z\in\mathscr{N}_h} w_z\varphi_z$ with $w_z = w_h(z) \in \mathbb{R}^m$ for every $z \in \mathscr{N}_h$, we thus have

$$\begin{aligned}\|\nabla w_h\|^2 &= \sum_{z,y\in\mathscr{N}_h} A_{zy} w_z \cdot w_y = \sum_{z,y\in\mathscr{N}_h} A_{zy}(w_z - w_y)\cdot w_y \\ &= \frac{1}{2}\sum_{z,y\in\mathscr{N}_h} A_{zy}(w_z - w_y)\cdot w_y + \frac{1}{2}\sum_{z,y\in\mathscr{N}_h} A_{zy}(w_y - w_z)\cdot w_z \\ &= -\frac{1}{2}\sum_{z,y\in\mathscr{N}_h} A_{zy}|w_z - w_y|^2.\end{aligned}$$

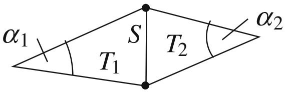


Fig. 3.7 Interior edge $S = T_1 \cap T_2$ with opposite angles α_1 and α_2

Therefore, the Lipschitz continuity of F and $A_{zy} \le 0$ for $z \neq y$ lead to

$$\begin{aligned}\|\nabla v_h^F\|^2 &= -\frac{1}{2}\sum_{z,y\in\mathscr{N}_h} A_{zy}\big|F(v_z)-F(v_y)\big|^2\\ &\le -\frac{1}{2}\sum_{z,y\in\mathscr{N}_h} A_{zy}\|DF\|^2_{L^\infty(\mathbb{R}^m)}|v_z-v_y|^2 = \|DF\|^2_{L^\infty(\mathbb{R}^m)}\|\nabla v_h\|^2,\end{aligned}$$

which proves the asserted estimate. □

Remarks 3.12 (i) If $d = 2$, then the conditions of the proposition are satisfied if and only if $\mathscr{T}_h$ is *weakly acute*, i.e., if every sum of two angles opposite to an interior edge is bounded by π and every angle opposite to an edge on the boundary by $\pi/2$. This follows from the relation

$$\int_{T_1\cup T_2} \nabla\varphi_z\cdot\nabla\varphi_y \,\mathrm{d}x = -\frac{1}{2}(\cot\alpha_1+\cot\alpha_2) = -\frac{1}{2}\frac{\sin(\alpha_1+\alpha_2)}{\sin(\alpha_1)\sin(\alpha_2)}$$

for neighboring triangles T_1, T_2 with common edge $S = \operatorname{conv}\{z, y\}$, cf. Fig. 3.7.
(ii) If $d = 3$, then a sufficient condition for the proposition is that every angle between two faces of a tetrahedron be bounded by $\pi/2$.
(iii) The conditions of the proposition imply that the finite element stiffness matrix $A = (A_{zy})_{z,y\in\mathscr{N}_h}$ is after elimination of rows and columns that correspond to nodes on Γ_D an M-matrix, i.e., that $Ax \ge 0$ implies $x \ge 0$ componentwise. This provides an alternative way to prove the discrete maximum principle.

Corollary 3.4 (Discrete maximum principle) *Assume that $\mathscr{T}_h$ is such that $\int_\Omega \nabla\varphi_z\cdot\nabla\varphi_y\,\mathrm{d}x \le 0$ for all distinct $z, y \in \mathscr{N}_h$. Then, if $u_h \in \mathscr{S}^1_\mathrm{D}(\mathscr{T}_h)$ satisfies $u_h(z) = u_\mathrm{D}(z)$ for all $z \in \mathscr{N}_h \cap \Gamma_\mathrm{D}$ and is minimal for*

$$I(u_h) = \frac{1}{2}\int_\Omega |\nabla u_h|^2\,\mathrm{d}x$$

in the set of all such functions in $\mathscr{S}^1(\mathscr{T}_h)$, then we have

$$\max_{z\in\mathscr{N}_h} u_h(z) \le \max_{z\in\mathscr{N}_h\cap\Gamma_\mathrm{D}} u_\mathrm{D}(z).$$

Proof Set $m_h = \max_{z\in\mathscr{N}_h\cap\Gamma_{\rm D}} u_{\rm D}(z)$ and note that the truncation operator $T_{m_h} : \mathbb{R}\to\mathbb{R}$, $s\mapsto\min\{s,m_h\}$ is Lipschitz continuous with constant $\|DT_{m_h}\|_{L^\infty(\mathbb{R})} = 1$. Thus, according to the previous proposition for $\widetilde{u}_h = \mathscr{I}_h(T_{m_h}\circ u_h)$, we have

$$I(\widetilde{u}_h) \le I(u_h),$$

which implies $\widetilde{u}_h = u_h$ and hence the asserted estimate. □

3.3 Approximation of the Heat Equation

We consider the linear heat equation which is for $f \in L^2([0,T];L^2(\Omega))$, $g \in L^\infty([0,T];L^2(\Gamma_{\rm N}))$, $\widetilde{u}_{\rm D} \in L^\infty([0,T];H^1(\Omega))$, and $u_0 \in L^2(\Omega)$ in a strong form for $t\in[0,T]$ defined by

$$\partial_t u - \Delta u = f \text{ in } \Omega, \quad u(0) = u_0, \quad u|_{\Gamma_{\rm D}} = \widetilde{u}_{\rm D}|_{\Gamma_{\rm D}}, \quad \partial_n u|_{\Gamma_{\rm N}} = g.$$

For simplicity, unless stated otherwise we restrict to the case $\widetilde{u}_{\rm D} = 0$ and $g = 0$. Throughout this section we use the notation

$$u_t = \partial_t u = u', \quad u_{tt} = \partial_t^2 u = u''.$$

Moreover, we abbreviate the inner product in $L^2(\Omega)$ by $(\cdot,\cdot)$ and use the Sobolev space $H^1_{\rm D}(\Omega)$ equipped with the norm $\|\nabla\cdot\|$. This requires $\Gamma_{\rm D}$ to have positive surface measure, but the results below can be generalized to the case $\Gamma_{\rm D} = \emptyset$. More general statements than the ones discussed below can be found in [12].

3.3.1 Variational Formulation

The discussion of gradient flows motivates the following definition of a weak solution of the heat equation.

Definition 3.12 A function $u \in H^1([0,T];H^1_{\rm D}(\Omega)')\cap L^2([0,T];H^1_{\rm D}(\Omega))$ is called a *weak solution of the heat equation* if $u(0) = u_0$ and

$$\langle u_t(t), v\rangle + (\nabla u(t), \nabla v) = (f(t), v)$$

for almost every $t\in[0,T]$ and all $v \in H^1_{\rm D}(\Omega)$.

Remark 3.13 If f is time-independent, then the heat equation is the L^2-gradient flow of the convex, Fréchet-differentiable functional $I : H^1_{\rm D}(\Omega)\to\mathbb{R}$ given by

$$I(u) = \frac{1}{2}\int_{\Omega} |\nabla u|^2 \,\mathrm{d}x - \int_{\Omega} f u \,\mathrm{d}x.$$

The discussion of subdifferential flows motivates that, for almost every $T' \in [0, T]$, the energy inequality

$$I(u(T')) + \int_0^{T'} \|\partial_t u(t)\|^2 \,\mathrm{d}t \leq I(u_0)$$

is satisfied. This implies that a weak solution is unique and belongs to the space $H^1([0, T]; L^2(\Omega)) \cap L^\infty([0, T]; H^1_{\mathrm{D}}(\Omega))$ provided that $I(u_0)$ is finite.

Theorem 3.7 (Existence and regularity) *There exists a unique weak solution u of the heat equation. If* $u_0 \in H^1_{\mathrm{D}}(\Omega)$, *then* $u \in H^1([0, T]; L^2(\Omega)) \cap L^\infty([0, T]; H^1_{\mathrm{D}}(\Omega))$. *If* Ω *is convex,* $\Gamma_{\mathrm{D}} = \partial\Omega$, $u_0 \in H^3(\Omega)$, $f \in H^1([0, T]; L^2(\Omega)) \cap L^2([0, T]; H^2(\Omega))$ *and* $u_0 \in H^1_0(\Omega)$, $u_1 = \Delta u_0 - f(0) \in H^1_0(\Omega)$, *then we have* $u_t \in L^2([0, T]; H^2(\Omega))$ *and* $u_{tt} \in L^2([0, T]; L^2(\Omega))$.

Proof (*sketched*) The first part of the theorem follows from the convexity of the Dirichlet energy. The second part exploits the H^2-regularity of the Laplace operator and a differentiation of the heat equation, i.e., considering $u'' - \Delta u' = f'$. □

Remark 3.14 The homogeneous heat equation with $f = 0$ has a regularizing effect, i.e., if $\partial\Omega$ is smooth, then for $u_0 \in L^2(\Omega)$, we have $u(t) \in C^\infty(\Omega)$ for every $t > 0$. On the other hand, we have $u(t) \to u_0$ in $L^2(\Omega)$ as $t \to 0$. Constructing smooth approximations of a function by mollification makes use of these properties. The regularizing effect is also reflected in the fact that the underlying diffusion process is irreversible. Mathematically, the time-reversed equation $u_t + \Delta u = 0$ in $[0, T] \times \Omega$ with $u(0) = u_0$ is ill-posed.

3.3.2 Semidiscrete in Time Approximation

We analyze various time-stepping schemes that will be the basis for fully discrete approximation schemes. Throughout the following, for any sequence $(a^k)_{k=0,\dots,K}$, we use the backward difference quotient defined by

$$d_t a^k = \frac{1}{\tau}(a^k - a^{k-1})$$

for $k = 1, 2, \dots, K$.

Lemma 3.8 (Difference calculus) *Given sequences* $(a^k)_{k=0,\dots,K}$ *and* $(b^k)_{k=0,\dots,K}$ *in a Hilbert space H, we have*

$$2(d_t a^k, a^k)_H = d_t \|a^k\|_H^2 + \tau \|d_t a^k\|_H^2.$$

Moreover, we have the discrete product rule $d_t(a^k, b^k)_H = (d_t a^k, b^k) + (a^{k-1}, d_t b^k)_H$ *and the* summation-by-parts *formula*

$$\tau \sum_{k=1}^{K} \Big((d_t a^k, b^k)_H + (a^{k-1}, d_t b^k)_H \Big) = (a^K, b^K)_H - (a^0, b^0)_H.$$

Proof The first identity follows from the binomial formula $2(a-b)a = (a^2 - b^2) + (a-b)^2$. The second identity is equivalent to $\tau d_t(a^k, b^k)_H = (a^k - a^{k-1}, b^k)_H + (a^{k-1}, b^k - b^{k-1})_H$, and the third identity follows from a summation over $k = 1, 2, \ldots, K$. □

The implicit Euler scheme leads to a sequence of equations that satisfy the conditions of the Lax–Milgram lemma.

Algorithm 3.1 (*Implicit Euler scheme*) Given $U^0 \in L^2(\Omega)$ and $\tau > 0$, compute for $k = 1, 2, \ldots, K$ with $K = \lfloor T/\tau \rfloor$ functions $U^k \in H^1_{\mathrm{D}}(\Omega)$ such that

$$(d_t U^k, v) + (\nabla U^k, \nabla v) = (f(t_k), v)$$

for all $v \in H^1_{\mathrm{D}}(\Omega)$.

To bound the error between the exact solution and the approximations $(U^k)_{k=0,\ldots,K}$, we first investigate consistency of the scheme.

Proposition 3.3 (Consistency) *Assume* $u \in C([0, T]; H^1_{\mathrm{D}}(\Omega))$ *and set* $u^k = u(t_k)$ *for* $k = 0, 1, \ldots, K$. *If* $u_{tt} \in L^2([0, T]; H^1_{\mathrm{D}}(\Omega)')$, *then we have*

$$(d_t u^k, v) + (\nabla u^k, \nabla v) = (f(t_k), v) + \mathscr{C}_\tau(t_k; v)$$

for all $v \in H^1_{\mathrm{D}}(\Omega)$ *with functionals* $\mathscr{C}_\tau(t_k) \in H^1_{\mathrm{D}}(\Omega)'$ *satisfying*

$$\tau \sum_{k=1}^{K} \|\mathscr{C}_\tau(t_k)\|^2_{H^1_{\mathrm{D}}(\Omega)'} \le c\tau^2.$$

Proof Due to the assumed regularity we have

$$\begin{aligned}(d_t u^k, v) + (\nabla u^k, \nabla v) &= (u_t(t_k), v) + (\nabla u(t_k), \nabla v) + (d_t u^k - u_t(t_k), v) \\ &= (f(t_k), v) + (d_t u^k - u_t(t_k), v)\end{aligned}$$

for all $v \in H^1_{\mathrm{D}}(\Omega)$. The identity

$$u_t(t_k) - d_t u^k = \frac{1}{\tau}\int_{t_{k-1}}^{t_k} \frac{d}{ds}\big((s-t_{k-1})u_t(s)\big) - u_t \,\mathrm{d}s = \frac{1}{\tau}\int_{t_{k-1}}^{t_k} (s-t_{k-1})u_{tt}\,\mathrm{d}s$$

implies that for every $v \in H^1_\mathrm{D}(\Omega)$ with $\|\nabla v\| \le 1$, we have

$$\begin{aligned}\mathscr{C}_\tau(t_k; v) = (d_t u^k - u_t(t_k), v) &= \frac{-1}{\tau}\int_{t_{k-1}}^{t_k} (s-t_{k-1})\langle u_{tt}, v\rangle\,\mathrm{d}s \\ &\le \Big(\tau^{-1}\int_{t_{k-1}}^{t_k} (s-t_{k-1})^2\,\mathrm{d}s\Big)^{1/2}\Big(\tau^{-1}\int_{t_{k-1}}^{t_k} \|u_{tt}\|^2_{H^1_\mathrm{D}(\Omega)'}\,\mathrm{d}s\Big)^{1/2}.\end{aligned}$$

We verify the estimate with $\int_{t_{k-1}}^{t_k}(s-t_{k-1})^2\,\mathrm{d}s = \tau^3/3$. □

Together with a discrete stability estimate, this implies a bound for the approximation error.

Proposition 3.4 (Discrete stability) *Suppose that the sequences $(z^k)_{k=0,\dots,K} \subset H^1_\mathrm{D}(\Omega)$ and $(b_k)_{k=1,\dots,K} \subset H^1_\mathrm{D}(\Omega)'$ satisfy*

$$(d_t z^k, v) + (\nabla z^k, \nabla v) = b_k(v)$$

for $k = 1, 2, \dots, K$. Then

$$\max_{k=1,\dots,K} \|z^k\|^2 + \tau\sum_{k=1}^K \|\nabla z^k\|^2 \le 2\|z_0\|^2 + 2\tau\sum_{k=1}^K \|b_k\|^2_{H^1_\mathrm{D}(\Omega)'}.$$

Proof Choosing $v = z^k$, we find with Lemma 3.8 that

$$\frac{d_t}{2}\|z^k\|^2 + \frac{\tau}{2}\|d_t z^k\|^2 + \|\nabla z^k\|^2 = b_k(z^k) \le \frac{1}{2}\|b_k\|^2_{H^1_\mathrm{D}(\Omega)'} + \frac{1}{2}\|\nabla z^k\|^2.$$

Multiplication by τ and summation over $k = 1, 2, \dots, L$ for $1 \le L \le K$ lead to

$$\|z^L\|^2 + \tau\sum_{k=1}^L \|\nabla z^k\|^2 \le \|z^0\|^2 + \sum_{k=1}^L \|b_k\|^2_{H^1_\mathrm{D}(\Omega)'}.$$

This proves the claimed estimate. □

The combination of the last two propositions implies the following error estimate.

Theorem 3.8 (Error estimate) *Under the assumptions of Proposition 3.3 we have*

$$\max_{k=1,2,\dots,K} \|u(t_k) - U^k\|^2 + \tau \sum_{k=1}^{K} \|\nabla(u(t_k) - U^k)\|^2 \le c\tau^2.$$

Proof The error $e^k = u^k - U^k$ satisfies $e^0 = 0$ and

$$(d_t e^k, v) + (\nabla e^k, \nabla v) = \mathscr{C}_\tau(t_k; v)$$

for $k = 1, 2, \dots, K$. The discrete stability estimate and the bound for the functionals $\mathscr{C}_\tau(t_k)$ lead to the estimate of the theorem. □

Remarks 3.15 (i) The assumption $u \in C([0,T]; H^1_D(\Omega))$ allows us to insert the sequence $(u^k)_{k=1,\dots,K}$ defined by $u^k = u(t_k)$ into the discrete scheme. Alternatively, one can employ the local temporal averages $u^k = (1/\tau)\int_{t_k-\tau/2}^{t_k+\tau/2} u(s)\,\mathrm{d}s$ for $k = 1, 2, \dots, K-1$ and $u^0 = u_0$.
(ii) By interpreting the heat equation as a gradient flow, a similar estimate can be proved under the sole condition that $-\Delta u_0 \in L^2(\Omega)$, cf. Theorem 4.7.

Under additional regularity assumptions, quadratic convergence with respect to τ can be proved for a modified scheme. Given sequences $(U^k)_{k=0,\dots,K}$ and $(t_k)_{k=0,\dots,K}$ we set for $k = 1, 2, \dots, K$

$$U^{k-1/2} = \frac{1}{2}(U^k + U^{k-1}), \quad t_{k-1/2} = \frac{1}{2}(t_k + t_{k-1}).$$

Algorithm 3.2 (*Crank–Nicolson scheme*) Given $U^0 \in H^1_D(\Omega)$ and $\tau > 0$, compute for $k = 1, 2, \dots, K$ with $K = \lfloor T/\tau \rfloor$ functions $U^k \in H^1_D(\Omega)$ such that

$$(d_t U^k, v) + (\nabla U^{k-1/2}, \nabla v) = (f(t_{k-1/2}), v)$$

for all $v \in H^1_D(\Omega)$.

Remark 3.16 If $u \in C^3([0,T] \times \overline{\Omega})$, then the Taylor expansions

$$u(t_k) = u(t_{k-1/2}) + (\tau/2)u_t(t_{k-1/2}) + (\tau^2/8)u_{tt}(t_{k-1/2}) + \mathscr{O}(\tau^3),$$
$$u(t_{k-1}) = u(t_{k-1/2}) - (\tau/2)u_t(t_{k-1/2}) + (\tau^2/8)u_{tt}(t_{k-1/2}) + \mathscr{O}(\tau^3)$$

show with $u^k = u(t_k)$, $k = 0, 1, \dots, K$, that

$$d_t u^k - u_t(t_{k-1/2}) = \mathscr{O}(\tau^2), \quad \nabla[u^{k-1/2} - u(t_{k-1/2})] = \mathscr{O}(\tau^2).$$

Therefore, we have

$$(d_t u^k, v) + (\nabla u^{k-1/2}, \nabla v) = (f(t_{k-1/2}), v) + \mathscr{C}^{cn}_\tau(t_{k-1/2}; v)$$

with

$$\mathscr{C}_\tau^{cn}(t_{k-1/2}; v) = (d_t u^k - u_t(t_{k-1/2}), v) + (\nabla[u^{k-1/2} - u(t_{k-1/2})], \nabla v)$$

which satisfies $\|\mathscr{C}_\tau^{cn}(t_{k-1/2})\|_{H^1_{\mathrm{D}}(\Omega)'} = \mathscr{O}(\tau^2)$.

The explicit Euler scheme is always unstable in the semidiscrete setting, e.g., it is undefined if $f = 0$ and $\Delta^k U^0 \notin L^2(\Omega)$ for some k with $1 \le k \le K$.

Algorithm 3.3 (*Explicit Euler scheme*) Given $U^0 \in H^1(\Omega)$ and $\tau > 0$, compute for $k = 1, 2, \ldots, K$ with $K = \lfloor T/\tau \rfloor$ functions $U^k \in H^1_{\mathrm{D}}(\Omega)$ such that

$$(d_t U^k, v) + (\nabla U^{k-1}, \nabla v) = (f(t_{k-1}), v)$$

for all $v \in H^1_{\mathrm{D}}(\Omega)$.

3.3.3 Semidiscrete in Space Approximation

To understand the influence of a spatial discretization of the heat equation, we consider the Galerkin method that defines a finite-dimensional system of ordinary differential equations.

Algorithm 3.4 (*Galerkin method*) Given a triangulation $\mathscr{T}_h$ of Ω and $u_{0,h} \in \mathscr{S}^1(\mathscr{T}_h)$, find $u_h \in H^1([0, T]; \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h))$ such that $u_h(0) = u_{0,h}$ and

$$(\partial_t u_h(t), v_h) + (\nabla u_h(t), \nabla v_h) = (f(t), v_h)$$

for almost every $t \in [0, T]$ and all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$.

We proceed as before and consider the consistency error for an interpolant of u. The obvious choice of the nodal interpolant $\mathscr{I}_h u(t)$ leads to

$$(\partial_t \mathscr{I}_h u, v_h) + (\nabla \mathscr{I}_h u, \nabla v_h) = (f(t), v_h) + \widetilde{\mathscr{C}_h}(t; v_h)$$

with

$$\widetilde{\mathscr{C}_h}(t; v_h) = (\partial_t[\mathscr{I}_h u - u], v_h) + (\nabla[\mathscr{I}_h u - u], \nabla v_h).$$

For a sufficiently regular solution u, the first term on the right-hand side is of order $\mathscr{O}(h^2)$, while the second term is only of order $\mathscr{O}(h)$. The alternative choice $Q_h u(t)$ with the H^1-projection of $u(t)$ onto $\mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ is known as *Wheeler's trick* and defines functionals $\mathscr{C}_h(t; \cdot)$ via

$$\mathscr{C}_h(t; v_h) = (\partial_t[Q_h u - u], v_h) + (\nabla[Q_h u - u], \nabla v_h) = (\partial_t[Q_h u - u], v_h).$$

Due to the Aubin–Nitsche lemma we have that $\mathscr{C}_h(t; v_h)$ is of order $\mathscr{O}(h^2)$. We make this observation precise in the following proposition.

Proposition 3.5 (Consistency) *If the Laplace operator is H^2-regular in Ω and $u_t \in L^2([0, T]; H^2(\Omega))$ then we have*

$$(\partial_t Q_h u, v_h) + (\nabla Q_h u, \nabla v_h) = (f, v_h) + \mathscr{C}_h(t; v_h)$$

with functionals $\mathscr{C}_h(t) \in H^1_\mathrm{D}(\Omega)'$ satisfying

$$\int_0^T \|\mathscr{C}_h\|^2_{H^1_\mathrm{D}(\Omega)'}\, \mathrm{d}t \le ch^4.$$

Proof The discussion above shows that it suffices to bound

$$\mathscr{C}_h(t; v_h) = (\partial_t(Q_h u - u), v_h).$$

Since Q_h is bounded and linear we have $\partial_t Q_h u = Q_h u_t$. With Theorem 3.5 we deduce that

$$\|Q_h u_t - u_t\| \le ch^2 \|D^2 u_t\|,$$

which holds for almost every $t \in [0, T]$ implies the result. □

An error estimate follows from a discrete stability estimate

Proposition 3.6 (Discrete stability) *Suppose that $z_h \in H^1([0, T]; \mathscr{S}^1_\mathrm{D}(\mathscr{T}_h))$ and $b_h \in L^2([0, T]; H^1_\mathrm{D}(\Omega)')$ satisfy*

$$(\partial_t z_h, v_h) + (\nabla z_h, \nabla v_h) = b_h(t; v_h)$$

for almost every $t \in [0, T]$ and every $v_h \in \mathscr{S}^1_\mathrm{D}(\mathscr{T}_h)$. Then

$$\sup_{t\in[0,T]} \|z_h(t)\|^2 + \int_0^T \|\nabla z_h\|^2\, \mathrm{d}t \le 2\|z_h(0)\|^2 + 2\int_0^T \|b_h\|^2_{H^1_\mathrm{D}(\Omega)'}\, \mathrm{d}t.$$

Proof The choice of $v_h = z_h(t)$ in the discrete equations immediately leads to the estimate. □

Theorem 3.9 (Error estimate) *Under the assumptions of Proposition 3.5 and if $\|u_{0,h} - Q_h u_0\| \le ch^2$, we have*

$$\sup_{t\in[0,T]} \|(u - u_h)(t)\|^2 + h^2 \int_0^T \|\nabla(u - u_h)\|^2\, \mathrm{d}t \le ch^4.$$

Proof The estimate for $u - u_h$ replaced by $Q_h u - u_h$ is a direct consequence of the consistency estimate and the discrete stability result. The triangle inequality and the estimates $\|u(t) - Q_h u(t)\| + h\|\nabla(u(t) - Q_h u(t))\| \le ch^2 \|D^2 u(t)\|$ then imply the bound for $u - u_h$. □

3.3.4 Fully Discrete Approximation

The explicit and implicit Euler scheme and the Crank–Nicolson scheme are special cases of the following θ-midpoint scheme related to $\theta = 0$, $\theta = 1$, and $\theta = 1/2$, respectively.

Algorithm 3.5 (*θ-midpoint scheme*) Given $\theta \in [0, 1]$, a triangulation $\mathscr{T}_h$ of Ω, and $u_h^0 \in \mathscr{S}^1(\mathscr{T}_h)$, compute for $k = 1, 2, \dots, K$ with $K = \lfloor T/\tau \rfloor$ functions $u_h^k \in \mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h)$ such that

$$(d_t u_h^k, v_h) + (\nabla[\theta u_h^k + (1-\theta)u_h^{k-1}], \nabla v_h) = (f(\theta t_k + (1-\theta)t_{k-1}), v_h)$$

for all $v_h \in \mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h)$.

We have unconditional stability if $\theta \ge 1/2$ and conditional stability if $\theta < 1/2$. We let $c_{\mathrm{inv}} > 0$ be such that $\|\nabla v_h\| \le c_{\mathrm{inv}} h^{-1} \|v_h\|$ for all $v_h \in \mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h)$ if $\mathscr{T}_h$ is quasiuniform.

Proposition 3.7 (Discrete stability) *Suppose that the sequences $(z_h^k)_{k=0,\dots,K} \subset \mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h)$ and $(b_k)_{k=0,\dots,K} \subset H_{\mathrm{D}}^1(\Omega)'$ satisfy*

$$(d_t z_h^k, v_h) + (\nabla[\theta z_h^k + (1-\theta)z_h^{k-1}], \nabla v_h) = b_k(v_h)$$

for all $v_h \in \mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h)$. If $\theta \ge 1/2$, we have

$$\max_{k=1,\dots,K} \|z_h^k\|^2 + \tau \sum_{k=1}^K \|\nabla[\theta z_h^k + (1-\theta)z_h^{k-1}]\|^2 \le 2\|z_h^0\|^2 + 2\tau \sum_{k=1}^K \|b_k\|_{H_{\mathrm{D}}^1(\Omega)'}^2.$$

Suppose that $\mathscr{T}_h$ is quasiuniform and $c_{\mathrm{inv}}^2 \tau h^{-2} \le 1/2$ if $\theta < 1/2$. Then

$$\max_{k=1,\dots,K} \frac{1}{2}\|z_h^k\|^2 + \tau \sum_{k=1}^K \|\nabla z_h^k\|^2 = 2\|z_h^0\|^2 + 2\tau \sum_{k=1}^K \|b_k\|_{H_{\mathrm{D}}^1(\Omega)'}^2.$$

Proof We abbreviate $z_h^{k,\theta} = \theta z_h^k + (1-\theta)z_h^{k-1}$ and assume first that $\theta \ge 1/2$. Noting $z_h^{k,\theta} = (z_h^k + z_h^{k-1})/2 + (\theta - 1/2)\tau d_t z_h^k$, the choice of $v_h = z_h^{k,\theta}$ yields

$$\frac{d_t}{2}\|z_h^k\|^2 + \big(\theta - \frac{1}{2}\big)\tau\|d_t z_h^k\|^2 + \|\nabla z_h^{k,\theta}\|^2 \le \frac{1}{2}\|b_k\|^2_{H^1_{\mathrm{D}}(\Omega)'} + \frac{1}{2}\|\nabla z_h^{k,\theta}\|^2.$$

A summation over $k = 1, 2, \dots, L$ for every $1 \le L \le K$ and multiplication by τ imply the estimate. If $\theta < 1/2$, then $v_h = z_h^k$ and $z_h^{k,\theta} = z_h^k - (1-\theta)\tau d_t z_h^k$ lead to

$$\frac{d_t}{2}\|z_h^k\|^2 + \frac{\tau}{2}\|d_t z_h^k\|^2 + \|\nabla z_h^k\|^2 = (1-\theta)\frac{\tau}{2}\big(d_t\|\nabla z_h^k\|^2 + \tau\|\nabla d_t z_h^k\|^2\big) + b_k(z_h^k).$$

Summing over $k = 1, 2, \dots, L$, multiplying by τ, and estimating $(1-\theta) \le 1$ show that

$$\frac{1}{2}\|z_h^L\|^2 + \frac{\tau^2}{2}\sum_{k=1}^{L}\|d_t z_h^k\|^2 + \frac{\tau}{2}\sum_{k=1}^{L}\|\nabla z_h^k\|^2 \le \frac{1}{2}\|z_h^0\|^2 + \frac{\tau}{2}\sum_{k=1}^{L}\|b_k\|^2_{H^1_{\mathrm{D}}(\Omega)'} + \frac{\tau}{2}\big(\|\nabla z_h^L\|^2 + \tau^2\sum_{k=1}^{L}\|\nabla d_t z_h^k\|^2\big).$$

We incorporate the inverse estimates

$$\|\nabla d_t z_h^k\| \le c_{\mathrm{inv}} h^{-1}\|d_t z_h^k\|, \quad \|\nabla z_h^L\| \le c_{\mathrm{inv}} h^{-1}\|z_h^L\|$$

to verify the stability estimate for $\theta < 1/2$. □

We verify the consistency of the numerical scheme only for the case $\theta = 1$.

Proposition 3.8 (Consistency) *If $u_{tt} \in L^2([0,T]; L^2(\Omega))$, $u_t \in L^2([0,T]; H^2(\Omega))$, $\theta = 1$, and the Poisson problem is H^2-regular, then we have for $u^k = u(t_k)$ that*

$$(d_t Q_h u^k, v_h) + (\nabla Q_h u^k, \nabla v_h) = (f(t_k), v_h) + \mathscr{C}_{h,\tau}(t_k; v_h)$$

with functionals $\mathscr{C}_{h,\tau}(t_k) \in H^1_{\mathrm{D}}(\Omega)'$ such that

$$\sum_{k=1}^{k}\|\mathscr{C}_{h,\tau}(t_k)\|^2_{H^1_{\mathrm{D}}(\Omega)'} \le c(\tau^2 + h^4).$$

Proof For $k = 1, 2, \dots, K$ we have, using $(\nabla(Q_h u^k - u^k), \nabla v_h) = 0$, that

$$\begin{aligned}(d_t Q_h u^k, v_h) + (\nabla Q_h u^k, \nabla v_h) &= (f(t_k), v_h) + (d_t Q_h u^k - \partial_t u(t_k), v_h)\\ &= (f(t_k), v_h) + \mathscr{C}_{h,\tau}(t_k; v_h).\end{aligned}$$

Arguing as in the proof of Proposition 3.3 and incorporating estimates for the H^1-projection, we find that the functional $\mathscr{C}_h(t_k; v)$ satisfies for every $v \in H^1_{\mathrm{D}}(\Omega)$ with $\|\nabla v\| \le 1$ the estimate

$$\begin{aligned}\mathscr{C}_{h,\tau}(t_k; v) &= (d_t Q_h u^k - d_t u^k, v) + (d_t u^k - \partial_t u(t_k), v)\\ &= \frac{1}{\tau}\int_{t_{k-1}}^{t_k} ((Q_h - 1)u_t, v)\,\mathrm{d}s + \frac{1}{\tau}\int_{t_{k-1}}^{t_k} (s - t_{k-1})(u_{tt}, v)\,\mathrm{d}s\\ &\le ch^2\Big(\int_{t_{k-1}}^{t_k} \|D^2 u_t\|^2\,\mathrm{d}t\Big)^{1/2} + c\tau\Big(\int_{t_{k-1}}^{t_k} \|u_{tt}\|^2\,\mathrm{d}t\Big)^{1/2}.\end{aligned}$$

This implies the asserted bound. □

Theorem 3.10 (Error estimate) *Under the condition of Proposition* 3.8 *we have*

$$\max_{k=1,\dots,K} \|u(t_k) - u_h^k\|^2 \le c(\tau^2 + h^4),$$

$$\tau\sum_{k=1}^{K} \|\nabla[u(t_k) - u_h^k]\|^2 \le c(\tau^2 + h^2).$$

Proof The estimate follows from the consistency result, the discrete stability, and the triangle inequality together with approximation properties of the H^1-projection. □

Remark 3.17 For the fully discrete Crank–Nicolson scheme corresponding to $\theta = 1/2$, one can prove the error bound $\max_{k=1,\dots,K} \|u(t_k) - u_h^k\| \le c(\tau^2 + h^2)$ under appropriate regularity conditions.

3.3.5 Discrete Maximum Principle

If $f \ge 0$ in $[0,T] \times \Omega$ and $u_0 \ge 0$, then the solution of the heat equation is nonnegative in $[0,T] \times \Omega$. Closely related is the (weak) maximum principle which states that if $f = 0$, then u attains its maximum on the boundary of $[0,T] \times \Omega$. For the semidiscrete scheme, which defines the approximation $U^k \in H^1_\mathrm{D}(\Omega)$ as the unique minimum of

$$I^k(U) = \frac{1}{2\tau}\|U - U^{k-1}\|^2 + \frac{1}{2}\int_\Omega |\nabla U|^2\,\mathrm{d}x,$$

we can argue by truncation as in the case of the Dirichlet energy. Letting $m^{k-1} = \max_{x\in\Omega} U^{k-1}(x)$ and setting

$$T_{m^{k-1}} U^k(x) = \min\{U^k(x), m^{k-1}\},$$

we have

$$I^k(T_{m^{k-1}}U^k) \le I^k(U^k)$$

and this implies $T_{m^{k-1}}U^k = U^k$, i.e., $U^k \le m^{k-1}$. An inductive argument implies that $\max_{k=0,\dots,K} \max_{x\in\Omega} U^k(x) \le \max_{x\in\Omega} U^0(x)$. We aim at using a similar argument in the fully discrete situation. This requires a modification of the numerical scheme.

Definition 3.13 Given a triangulation $\mathscr{T}_h$ of Ω, the *discrete* (*or lumped*) L^2*-inner product* is for $v, w \in C(\overline{\Omega})$ defined as

$$(v, w)_h = \int_\Omega \mathscr{I}_h(vw)\,\mathrm{d}x = \sum_{z\in\mathscr{N}_h} \beta_z v(z)w(z)$$

with $\beta_z = \int_\Omega \varphi_z\,\mathrm{d}x$ for every $z \in \mathscr{N}_h$. The corresponding *discrete* (*or lumped*) (*semi-*) *norm* is for $v \in C(\overline{\Omega})$ defined by $\|v\|_h^2 = (v, v)_h$.

Lemma 3.9 (Discrete inner product) *For $v_h, w_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ we have*

$$\|v_h\| \le \|v_h\|_h \le (d+2)^{1/2}\|v_h\|$$

and

$$|(v_h, w_h)_h - (v_h, w_h)| \le ch^{1+\ell}\|\nabla v_h\|\|\nabla^\ell w_h\|$$

for $\ell \in \{0, 1\}$ and $\nabla^0 w_h = w_h$ and $\nabla^1 w_h = \nabla w_h$.

Proof For every $T \in \mathscr{T}_h$ such that $T = \mathrm{conv}\{z_0, z_1, \dots, z_d\}$ with $z_0, z_1, \dots, z_d \in T \cap \mathscr{N}_h$ a transformation argument shows for $M_T = |T|d!/(d+2)!$ that

$$\int_T \varphi_{z_j}\varphi_{z_k}\,\mathrm{d}x = (1+\delta_{jk})M_T, \quad \int_T \mathscr{I}_h(\varphi_{z_j}\varphi_{z_k})\,\mathrm{d}x = (d+2)\delta_{jk}M_T.$$

With these identities it follows that for $v_h|_T = \sum_{j=0}^d a_j\varphi_{z_j}$, we have

$$M_T^{-1}\|v_h\|^2_{L^2(T)} = 2\sum_{j=0}^d a_j^2 + \sum_{\substack{j,k=0\\ j\ne k}}^d a_j a_k \le (d+2)\sum_{j=0}^d a_j^2 = M_T^{-1}\|v_h\|^2_{h,T},$$

where we abbreviated $\int_T \mathscr{I}_h(v_h^2)\,\mathrm{d}x = \|v_h\|^2_{h,T}$. Conversely, we have

$$M_T^{-1}\|v_h\|^2_{h,T} \le 2(d+2)\sum_{j=0}^d a_j^2 + (d+2)\sum_{\substack{j,k=0\\ j\ne k}}^d a_j a_k = (d+2)M_T^{-1}\|v_h\|^2_{L^2(T)}.$$

The estimates for nodal interpolation, together with an inverse inequality if $d \geq 3$ and $D^2 v_h|_T = 0$, $D^2 w_h|_T = 0$, show that

$$\int_T \big|v_h w_h - \mathscr{I}_h(v_h w_h)\big| \, \mathrm{d}x \leq c h_T^2 \int_T |D^2(v_h w_h)| \, \mathrm{d}x \leq c h_T^2 \|\nabla v_h\|_{L^2(T)} \|\nabla w_h\|_{L^2(T)}.$$

A summation over all $T \in \mathscr{T}_h$ proves the estimates for $\ell = 1$. To prove the estimate for the case $\ell = 0$, we first employ the inverse estimate $\|\nabla w_h\|_{L^2(T)} \leq c h_T^{-1} \|w_h\|_{L^2(T)}$ for every $T \in \mathscr{T}_h$. □

Remarks 3.18 (i) The discrete inner product has a stabilizing effect, e.g., for $d = 1$ we have for all $v_h \in \mathscr{S}^1(\mathscr{T}_h)$ that

$$\|v_h\|_h^2 = \|v_h\|^2 + \frac{1}{6} \sum_{T \in \mathscr{T}_h} h_T^2 \|\nabla v_h\|_{L^2(T)}^2.$$

(ii) The discrete inner product allows for a localization of quantities since $(v, \varphi_z)_h = \beta_z v(z)$ for every $v \in C(\overline{\Omega})$ and every $z \in \mathscr{N}_h$.
(iii) The use of the discrete inner product is also referred to as *reduced integration*.

A discrete maximum principle holds for the modified implicit Euler scheme which employs the discrete L^2-inner product.

Theorem 3.11 (Discrete maximum principle) *Assume that $\mathscr{T}_h$ is weakly acute and $f \in C([0, T] \times \overline{\Omega})$. Let $(u_h^k)_{k=0,\dots,K} \in \mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h)$ satisfy*

$$(d_t u_h^k, v_h)_h + (\nabla u_h^k, \nabla v_h) = (f(t_k), v_h)_h$$

for $k = 1, 2, \dots, K$ and all $v_h \in \mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h)$. If $f \geq 0$ in $[0, T] \times \Omega$ and $u_h^0 \geq 0$ in Ω, then $u_h^k \geq 0$ in Ω for $k = 1, 2, \dots, K$. If $f = 0$, then

$$\max_{k=1,\dots,K} \max_{z \in \mathscr{N}_h} u_h^k(z) \leq \max_{z \in \mathscr{N}_h} u_h^0(z).$$

Proof Assume that for $1 \leq k \leq K$ we have $u_h^{k-1} \geq 0$ and $f(t_k) \geq 0$. The function u_h^k is the unique minimizer of the functional

$$I_h^k(u_h) = \frac{1}{2\tau} \|u_h - u_h^{k-1}\|_h^2 + \frac{1}{2} \int_\Omega |\nabla u_h|^2 \, \mathrm{d}x - \int_\Omega \mathscr{I}_h[f u_h] \, \mathrm{d}x$$

in the set of functions $u_h \in \mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h)$. We define the function $\widehat{u}_h^k \in \mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h)$ through the truncated nodal values $\widehat{u}_h^k(z) = \max\{u_h^k(z), 0\}$ for all $z \in \mathscr{N}_h$. Then $\widehat{u}_h^k \geq 0$ in Ω and

$$|\widehat{u}_h^k(z) - u_h^{k-1}(z)| \leq |u_h^k(z) - u_h^{k-1}(z)|$$

for all $z \in \mathcal{N}_h$. Arguing as in the proof of Corollary 3.4, we find $I_h^k(\widehat{u}_h^k) \le I_h^k(u_h^k)$. This implies $u_h^k = \widehat{u}_h^k$ and $u_h^k \ge 0$. If $f = 0$, we set $m_h^{k-1} = \max_{z \in \mathcal{N}_h} u_h^{k-1}(z)$ and define the function $\widetilde{u}_h^k \in \mathcal{S}_{\mathrm{D}}^1(\mathcal{T}_h)$ through the truncated nodal values

$$\widetilde{u}_h^k(z) = \min\{u_h^k(z), m_h^{k-1}\}.$$

Again, we have $|\widetilde{u}_h^k(z) - u_h^{k-1}(z)| \le |u_h^k(z) - u_h^{k-1}(z)|$ for all $z \in \mathcal{N}_h$ and $I_h^k(\widetilde{u}_h^k) \le I_h^k(u_h^k)$. Therefore $u_h^k = \widetilde{u}_h^k \le m_h^{k-1}$ and an inductive argument finishes the proof. □

Remarks 3.19 (i) The nonnegativity of the solutions can also be proved by noting that the nontrivial nodal values $\widehat{U}^k = (u_h^k(z))_{z \in \mathcal{N}_h \setminus \Gamma_{\mathrm{D}}}$ solve the linear systems of equations

$$(\widehat{M}_h + \tau \widehat{A})\widehat{U}^k = \widehat{M}_h \widehat{U}^{k-1} + \tau \widehat{M}_h \widehat{F}^k$$

with the diagonal mass matrix $\widehat{M}_h$ related to the discrete inner product $(\cdot, \cdot)_h$ and the finite element stiffness matrix $\widehat{A}$. The matrix $\widehat{A}$ is diagonally dominant and has positive entries only on the diagonal because of the weak acuteness of the underlying triangulation. Therefore, the matrix $\widehat{M}_h + \tau \widehat{A}$ is an M-matrix and its inverse has nonnegative entries.
(ii) Approximations obtained with the Crank–Nicolson scheme do in general not satisfy a maximum principle even if the discrete inner product is used.

3.3.6 A Posteriori Error Estimate

The schemes discussed above can easily be modified to allow for variable time steps $(\tau_k)_{k=1,\dots,K}$ and triangulations $(\mathcal{T}_h^k)_{k=0,\dots,K}$. We then set $t_k = \sum_{j=1}^k \tau_j$, $k = 0, 1, \dots, K$, assume that $t_K = T$, and define

$$d_t u_h^k = \frac{1}{\tau_k}(u_h^k - u_h^{k-1}), \qquad \widetilde{d}_t u_h^k = \frac{1}{\tau_k}(u_h^k - \mathcal{I}_h^k u_h^{k-1}),$$

with the nodal interpolant $\mathcal{I}_h^k$ associated with $\mathcal{S}_{\mathrm{D}}^1(\mathcal{T}_h^k)$. For a sequence of approximations $(u_h^k)_{k=0,\dots,K}$ related to the time steps $(t_k)_{k=0,\dots,K}$ and such that $u_h^k \in \mathcal{S}_{\mathrm{D}}^1(\mathcal{T}_h^k)$, we define the continuous interpolant

$$\widehat{u}_{h,\tau}(t,x) = \frac{t - t_{k-1}}{\tau_k} u_h^k(x) + \frac{t_k - t}{\tau_k} u_h^{k-1}(x)$$

for $x \in \Omega$ and $t \in [t_{k-1}, t_k]$ and $k = 1, 2, \dots, K$.

Proposition 3.9 (Residual estimate) *Assume that $f \in C^1([0,T]; L^2(\Omega))$. For $k = 0, 1, \dots, K$, let $\mathcal{T}_h^k$ be a triangulation of Ω and assume that $u_h^k \in \mathcal{S}_{\mathrm{D}}^1(\mathcal{T}_h^k)$ satisfies*

$$(\widetilde{d}_t u_h^k, v_h) + (\nabla u_h^k, \nabla v_h) = (f(t_k), v_h)$$

for $k = 1, 2, \ldots, K$ *and all* $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h^k)$. *For almost every* $t \in [0, T]$, *define* $\mathscr{R}_{h,\tau}(t) \in H^1_{\mathrm{D}}(\Omega)'$ *by*

$$\mathscr{R}_{h,\tau}(t; v) = (\partial_t \widehat{u}_{h,\tau}(t), v) + (\nabla \widehat{u}_{h,\tau}(t), \nabla v) - (f(t), v)$$

for every $v \in H^1_{\mathrm{D}}(\Omega)$. *For* $k = 1, 2, \ldots, K$ *and almost every* $t \in [t_{k-1}, t_k]$, *we have*

$$\|\mathscr{R}_{h,\tau}(t)\|^2_{H^1_{\mathrm{D}}(\Omega)'} \le c\big(\eta^2_{\text{space}}(t_k) + \eta^2_{\text{time}}(t_k) + \eta^2_{\text{coarse}}(t_k) + \eta^2_{\text{data}}(t_k)\big)$$

with the space discretization residual

$$\eta^2_{\text{space}}(t_k) = \sum_{T \in \mathscr{T}_h^k} h_T^2 \|\widetilde{d}_t u_h^k - \Delta u_h^k - f(t_k)\|^2_{L^2(T)} \\ + \sum_{S \in \mathscr{S}_h^k \cap \Omega} h_S \|[\![\nabla u_h^k \cdot n_S]\!]\|^2_{L^2(S)} + \sum_{S \in \mathscr{S}_h^k \cap \overline{\Gamma}_{\mathrm{N}}} h_S \|\partial_n u_h^k\|^2_{L^2(S)},$$

the time discretization residual

$$\eta^2_{\text{time}}(t_k) = \|\nabla[u_h^{k-1} - u_h^k]\|^2,$$

the mesh coarsening residual

$$\eta^2_{\text{coarse}}(t_k) = \tau_k^{-2} \|\mathscr{I}_h^k u_h^{k-1} - u_h^{k-1}\|^2,$$

and the data approximation residual

$$\eta^2_{\text{data}}(t_k) = \tau_k^2 \sup_{t \in [t_{k-1}, t_k]} \|\partial_t f(t)\|^2$$

Proof Let $t \in (t_{k-1}, t_k)$ and $v \in H^1_{\mathrm{D}}(\Omega)$. Then $\partial_t \widehat{u}_{h,\tau}(t) = d_t u_h^k$ and we have

$$\begin{aligned} \mathscr{R}_{h,\tau}(t; v) &= (\widetilde{d}_t u_h^k, v) + (\nabla u_h^k, \nabla v) - (f(t_k), v) \\ &\quad + (\nabla[\widehat{u}_{h,\tau}(t) - u_h^k], \nabla v) - (f(t) - f(t_k), v) + (\widetilde{d}_t u_h^k - d_t u_h^k, v) \\ &= (d_t u_h^k, v - v_h) + (\nabla u_h^k, \nabla[v - v_h]) - (f(t_k), v - v_h) \\ &\quad + (\nabla[\widehat{u}_{h,\tau}(t) - u_h^k], \nabla v) - (f(t) - f(t_k), v) \\ &\quad + \frac{1}{\tau_k}(\mathscr{I}_h^k u_h^{k-1} - u_h^{k-1}, v) \\ &= I + II + \ldots + VI. \end{aligned}$$

With $v_h = \mathscr{J}_h^k v \in \mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h^k)$ and an elementwise integration-by-parts as in the proof of Theorem 3.6, the first three terms $I + II + III$ on the right-hand side satisfy

$$I + II + III \le c\,\eta_{\mathrm{space}}^2(t_k)\|\nabla v\|.$$

The terms $IV + V + VI$ are estimated with Hölder and Poincaré inequalities. □

Proposition 3.10 (Continuous stability) *Assume that* $z \in H^1([0,T]; H_{\mathrm{D}}^1(\Omega)') \cap L^2([0,T]; H^1(\Omega))$ *and* $b \in L^2([0,T]; H_{\mathrm{D}}^1(\Omega)')$ *satisfy*

$$\langle \partial_t z, v\rangle + (\nabla z, \nabla v) = b(v)$$

for almost every $t \in [0,T]$ *and every* $v \in H_{\mathrm{D}}^1(\Omega)$. *We then have*

$$\sup_{t\in[0,T]} \|z(t)\|^2 + \int_0^T \|\nabla z\|^2 \,\mathrm{d}t \le 2\|z(0)\|^2 + 2\int_0^T \|b\|_{H_{\mathrm{D}}^1(\Omega)'}^2 \,\mathrm{d}t.$$

Proof The proof follows from choosing $v = z(t)$ and integrating the resulting identity over $t \in [0,T]$. □

Theorem 3.12 (A posteriori error estimate) *Under the conditions of Proposition 3.9, we have*

$$\begin{aligned}\sup_{t\in[0,T]} \|(u-\widehat{u}_{h,\tau})(t)\|^2 + \int_0^T \|\nabla(u-\widehat{u}_{h,\tau})\|^2\,\mathrm{d}t &\le 2\|u_0 - \widehat{u}_{h,\tau}(0)\|^2 \\ &\quad + 2c\sum_{k=1}^K \tau_k\big(\eta_{\mathrm{space}}^2(t_k) + \eta_{\mathrm{time}}^2(t_k) \\ &\quad + \eta_{\mathrm{coarse}}^2(t_k) + \eta_{\mathrm{data}}^2(t_k)\big).\end{aligned}$$

Proof The estimate follows from a straightforward combination of the residual bound and the continuous stability estimate. □

Remark 3.20 The theorem provides a computable upper bound for the approximation error. Since it is the sum of local quantities, it can be used to refine and coarsen the mesh-size and the time-steps locally.

3.4 Implementation of the $P1$ Finite Element Method

We describe in this section a way of implementing the $P1$ finite element method. Several ideas reported below are adopted from [1, 4].

3.4.1 Poisson Problem

We make the following assumption on the data functions in order to assume an exact and simple numerical integration. The influence of the approximation of possibly discontinuous data functions by such functions can be analyzed within the Strang lemmas.

Assumption 3.1 (*Data approximation I*) We assume that $u_{\mathrm{D}} = \widetilde{u}_{\mathrm{D},h}|_{\Gamma_{\mathrm{D}}}$ for a function $\widetilde{u}_{\mathrm{D},h} \in \mathscr{S}^1(\mathscr{T}_h)$ and $f = f_h$ and $g = g_h$ are piecewise constant.

We write the unknown function as $u_h = \widetilde{u}_h + \widetilde{u}_{\mathrm{D},h}$ and seek the uniquely defined function $\widetilde{u}_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ that satisfies

$$\int_\Omega \nabla\widetilde{u}_h \cdot \nabla v_h \,\mathrm{d}x = \int_\Omega f_h v_h \,\mathrm{d}x + \int_{\Gamma_{\mathrm{N}}} g_h v_h \,\mathrm{d}s - \int_\Omega \nabla\widetilde{u}_{\mathrm{D},h} \cdot \nabla v_h \,\mathrm{d}x$$

for all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$. We let $\widetilde{U} = (\widetilde{U}_y : y \in \mathscr{K}_h)$ be the coefficients of $\widetilde{u}_h$ with respect to the nodal basis restricted to the*free nodes* $\mathscr{K}_h = \mathscr{N}_h \setminus \Gamma_{\mathrm{D}}$. For every $T \in \mathscr{T}_h$ and every $S \in \mathscr{S}_h$ we let $x_T = (1/(d+1)) \sum_{z \in \mathscr{N}_h \cap T} z$ and $x_S = (1/d) \sum_{z \in \mathscr{N}_h \cap S} z$ denote their midpoints and note that the corresponding one-point quadrature rules are exact for affine functions. The discrete formulation is thus equivalent to the linear system of equations

$$\begin{aligned}\sum_{y \in \mathscr{K}_h} \widetilde{U}_y \int_\Omega \nabla\varphi_y \cdot \nabla\varphi_z \,\mathrm{d}x = \sum_{T \in \mathscr{T}_h} f_h(x_T) \int_T \varphi_z \,\mathrm{d}x + \sum_{S \subset \mathscr{S}_h \cap \overline{\Gamma}_{\mathrm{N}}} g_h(x_S) \int_S \varphi_z \,\mathrm{d}s \\ - \sum_{y \in \mathscr{N}_h} \widetilde{u}_{\mathrm{D},h}(y) \int_\Omega \nabla\varphi_y \cdot \nabla\varphi_z \,\mathrm{d}x\end{aligned}$$

for all $z \in \mathscr{K}_h$, i.e., $\widetilde{s}\widetilde{U} = \widetilde{b}$ with a symmetric matrix $\widetilde{s} \in \mathbb{R}^{\#\mathscr{K}_h \times \#\mathscr{K}_h}$ and $\widetilde{b} \in \mathbb{R}^{\#\mathscr{K}_h}$. The integrals that define the matrix and the vector on the right-hand side are computed by decomposing the integral as a sum over elements, e.g.,

$$\int_\Omega \nabla\varphi_z \cdot \nabla\varphi_y \,\mathrm{d}x = \sum_{T \in \mathscr{T}_h :\, z,y \in T} \int_T \nabla\varphi_z \cdot \nabla\varphi_y \,\mathrm{d}x.$$

The triangulation of Ω and the partition of the boundary $\partial\Omega$ are defined through the arrays c4n, n4e, Db, and Nb that specify the coordinates of the nodes, the vertices of the elements, and the vertices of the sides on Γ_{D} and $\overline{\Gamma}_{\mathrm{N}}$, respectively. In particular, the $n_C \times d$ array c4n defines the coordinates of the nodes and implicitly an enumeration of the nodes. The $n_E \times (d+1)$ array n4e defines the elements by specifying the positions of their vertices through their numbers. Similarly, the $n_{Db} \times d$ and $n_{Nb} \times d$ arrays Db and Nb define the vertices of the sides belonging

$\Omega = (0,1)^2$, $\Gamma_{\mathrm{D}} = \{0,1\} \times \{0\}$, $\Gamma_{\mathrm{N}} = \partial\Omega \setminus \Gamma_{\mathrm{D}}$

```
c4n = [0,0;1,0;1,1;0,1];
n4e = [1,2,3;1,3,4];
Db = [1,2];
Nb = [2,3;3,4;4,1];
```

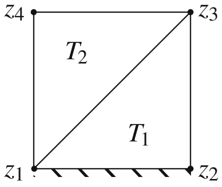


Fig. 3.8 Triangulation of the unit square and corresponding arrays

to Γ_{D} and $\overline{\Gamma}_{\mathrm{N}}$, respectively. The arrays are displayed in Fig. 3.8 for a triangulation consisting of two triangles and with four nodes.

Assumption 3.2 (*Orientation*) We assume that the list of elements defines an ordering of the nodes of elements that induces a positive orientation of T, i.e., if $T \equiv (z_0, z_1, \dots, z_d)$ for $T \in \mathscr{T}_h$ and $z_0, z_1, \dots, z_d \in \mathscr{N}_h$ such that $T = \mathrm{conv}\{z_0, z_1, \dots, z_d\}$, then the vectors $\tau_\ell = z_\ell - z_0$, $\ell = 1, 2, \dots, d$, satisfy

$$\tau_1 > 0, \quad \tau_2 \cdot \tau_1^{\perp} > 0, \quad \tau_3 \cdot (\tau_1 \times \tau_2) > 0$$

for $d = 1, 2, 3$, respectively.

To compute the system matrix $\widetilde{s}$ and the vector $\widetilde{b}$ on the right-hand side of the linear system of equations stated above, we note some elementary identities for the nodal basis functions.

Lemma 3.10 (Elementwise gradients) *Let* $T \equiv (z_0, z_1, \dots, z_d)$ *with* $z_0, z_1, \dots, z_d \in \mathbb{R}^d$ *and define*

$$X_T = \begin{bmatrix} 1 & 1 & \dots & 1 \\ z_0 & z_1 & \dots & z_d \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}.$$

We then have that the volume $|T|$ *is given by* $|T| = (1/d!) \det X_T$, *and with the identity matrix* $I_d \in \mathbb{R}^{d\times d}$ *that*

$$\left[\nabla\varphi_{z_0}|_T, \dots, \nabla\varphi_{z_d}|_T\right]^\top = X_T^{-1} \begin{bmatrix} 0 \\ I_d \end{bmatrix}.$$

Proof The proof follows from noting that the nodal basis function associated to z_j is for $x \in T$ given by

$$\varphi_{z_j}(x) = \frac{1}{d!|T|} \det \begin{bmatrix} 1 & 1 & \dots & 1 \\ x & z_{j+1} & \dots & z_{j+d} \end{bmatrix},$$

where subscripts are understood modulo d, together with Laplace's formula and Cramer's rule. □

Some additional identities are required for computing the vector $\widetilde{b}$.

Lemma 3.11 (Right-hand side) *For a side $S = \operatorname{conv}\{z_0, z_1, \dots, z_{d-1}\} \in \mathscr{S}_h$, the surface area $|S|$ is given by*

$$|S| = \begin{cases} 1 & \text{if } d = 1, \\ |z_1 - z_0| & \text{if } d = 2, \\ |(z_2 - z_0) \times (z_1 - z_0)|/2 & \text{if } d = 3. \end{cases}$$

Moreover, for $T \in \mathscr{T}_h$, $S \in \mathscr{S}_h$, and $z \in T \cap S$, we have

$$\int_T \varphi_z \, dx = \frac{|T|}{d+1}, \quad \int_S \varphi_z \, ds = \frac{|S|}{d}.$$

Proof The proof of the formula for $|S|$ follows from elementary geometric identities. The integrals over T and S are computed with the help of an affine transformation to a reference element. □

Figure 3.9 shows a MATLAB implementation of the $P1$ method in which the matrix $\widetilde{s}$ corresponds to the array `s(fNodes,fNodes)`. We input the space dimension and the number of refinements of a coarse triangulation. The routine `red_refine.m` carries out the refinements of the triangulation by dividing every element into 2^d subelements. The operation `s`β solves a linear system of equations and the command `sparse(I,J,X,nC,nC)` assembles a sparse matrix $s \in \mathbb{R}^{nc \times nc}$ by specifying its entries through lists $I, J, X \in \mathbb{R}^L$ and $s_{ij} = \sum_{\ell \in \{1,\dots,L\}, I_\ell = i, J_\ell = j} X_\ell$.

3.4.2 Heat Equation

For a simple implementation, different assumptions on the approximation of the data functions are made for the implementation of the θ-midpoint scheme for the heat equation.

Assumption 3.3 (*Data approximation II*) We assume that

$$u_0 = u_{0,h} \in \mathscr{S}^1(\mathscr{T}_h), \qquad u_{\mathrm{D}} = u_{\mathrm{D},h} \in C([0,T]; \mathscr{S}^1(\mathscr{T}_h)),$$
$$f = f_h \in C([0,T]; \mathscr{S}^1(\mathscr{T}_h)), \quad g = g_h \in C([0,T]; \mathscr{S}^1(\mathscr{T}_h)).$$

For a sequence $(a^k)_{k=0,\dots,K}$ and $\theta \in [0,1]$, we set $a^{k,\theta} = \theta a^k + (1-\theta)a^{k-1}$. The θ-midpoint scheme then computes the sequence $(\widetilde{u}_h^k)_{k=0,\dots,K} \subset \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ with

$$\int_\Omega d_t \widetilde{u}_h^k v_h \, dx + \int_\Omega \nabla \widetilde{u}_h^{k,\theta} \cdot \nabla v_h \, dx = -\int_\Omega d_t u_{\mathrm{D},h}^k v_h \, dx - \int_\Omega \nabla u_{\mathrm{D},h}^{k,\theta} \cdot \nabla v_h \, dx$$
$$+ \int_\Omega f_h(t_{k,\theta}) v_h \, dx + \int_{\Gamma_{\mathrm{N}}} g_h(t_{k,\theta}) v_h \, ds$$

```
function p1_poisson(d,red)
[c4n,n4e,Db,Nb] = triang_cube(d);
for j = 1:red
    [c4n,n4e,Db,Nb,Pr0,Pr1] = red_refine(c4n,n4e,Db,Nb);
end
[nC,d] = size(c4n); nE = size(n4e,1); nNb = size(Nb,1);
dNodes = unique(Db); fNodes = setdiff(1:nC,dNodes);
u = zeros(nC,1); tu_D = zeros(nC,1); b = zeros(nC,1);
ctr = 0; ctr_max = (d+1)^2*nE;
I = zeros(ctr_max,1); J = zeros(ctr_max,1); X = zeros(ctr_max,1);
for j = 1:nE
    X_T = [ones(1,d+1);c4n(n4e(j,:),:)'];
    grads_T = X_T\[zeros(1,d);eye(d)];
    vol_T = det(X_T)/factorial(d);
    mp_T = sum(c4n(n4e(j,:),:),1)/(d+1);
    for m = 1:d+1
        b(n4e(j,m)) = b(n4e(j,m))+(1/(d+1))*vol_T*f(mp_T);
        for n = 1:d+1
            ctr = ctr+1; I(ctr) = n4e(j,m); J(ctr) = n4e(j,n);
            X(ctr) = vol_T*grads_T(m,:)*grads_T(n,:)';
        end
    end
end
s = sparse(I,J,X,nC,nC);
for j = 1:nNb
    if d == 1
        vol_S = 1;
    elseif d == 2
        vol_S = norm(c4n(Nb(j,1),:)-c4n(Nb(j,2),:));
    elseif d == 3
        vol_S = norm(cross(c4n(Nb(j,3),:)-c4n(Nb(j,1),:),...
            c4n(Nb(j,2),:)-c4n(Nb(j,1),:)),2)/2;
    end
    mp_S = sum(c4n(Nb(j,:),:),1)/d;
    for k = 1:d
        b(Nb(j,k)) = b(Nb(j,k))+(1/d)*vol_S*g(mp_S);
    end
end
for j = 1:nC
    tu_D(j) = u_D(c4n(j,:));
end
b = b-s*tu_D; u(fNodes) = s(fNodes,fNodes)\b(fNodes); u = u+tu_D;
if d == 1
    plot(c4n(n4e),u(n4e));
elseif d == 2
    trisurf(n4e,c4n(:,1),c4n(:,2),u);
elseif d == 3
    trisurf([Db;Nb],c4n(:,1),c4n(:,2),c4n(:,3),u);
end

function val = f(x); val = 1;
function val = g(x); val = 1;
function val = u_D(x); val = sin(2*pi*x(:,1));
```

Fig. 3.9 MATLAB implementation of the $P1$ finite element method for the Poisson problem

```
function p1_theta_heat(d,red)
T = 10; theta = 1/2; alpha = 1;
[c4n,n4e,Db,Nb] = triang_cube(d);
for j = 1:red
    [c4n,n4e,Db,Nb,Pr0,Pr1] = red_refine(c4n,n4e,Db,Nb);
end
nC = size(c4n,1);
dNodes = unique(Db); fNodes = setdiff(1:nC,dNodes);
h = 2^(-red); tau = h^alpha/4; K = floor(T/tau);
u_old = u_0(c4n)-u_D(0,c4n); u_new = zeros(nC,1);
[s,m,m_lumped,vol_T] = fe_matrices(c4n,n4e);
[m_Nb,m_Nb_lumped] = fe_matrices_bdy(c4n,Nb);
for k = 1:K
    t_k = k*tau; t_k_theta = (k-1+theta)*tau;
    dt_u_D = (1/tau)*(u_D(t_k,c4n)-u_D(t_k-tau,c4n));
    u_D_k_theta = theta*u_D(t_k,c4n)+(1-theta)*u_D(t_k-tau,c4n);
    b = (1/tau)*m*u_old-m*dt_u_D...
        -s*u_D_k_theta-(1-theta)*s*u_old...
        +m*f(t_k_theta,c4n)+m_Nb*g(t_k_theta,c4n);
    X = (1/tau)*m+theta*s;
    u_new(fNodes) = X(fNodes,fNodes)\b(fNodes);
    show_p1(c4n,n4e,Db,Nb,u_new+u_D(t_k,c4n)); drawnow;
    u_old = u_new;
end

function val = f(t,x); val = ones(size(x,1),1);
function val = u_0(x); val = sin(2*pi*x(:,1));
function val = u_D(t,x); val = min(t,.2)*sin(2*pi*x(:,1));
function val = g(t,x); val = zeros(size(x,1),1);
```

Fig. 3.10 MATLAB implementation of the θ-midpoint scheme in time and the $P1$ finite element method in space for the heat equation

for all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$. We also set $u^k_{\mathrm{D},h} = u_{\mathrm{D}}(t_k)$ for $k = 0, 1, \ldots, K$. The nontrivial coefficients $\widetilde{U}^k = (\widetilde{U}^k_y : y \in \mathscr{K}_h)$ of $\widetilde{u}^k_h$ satisfy the equation

$$\begin{aligned}
\sum_{y\in\mathscr{K}_h} \widetilde{U}^k_y \Big(\tau^{-1} \int_\Omega \varphi_z\varphi_y \,\mathrm{d}x + \theta \int_\Omega \nabla\varphi_y \cdot \nabla\varphi_z \,\mathrm{d}x\Big)& \\
= \sum_{y\in\mathscr{K}_h} \widetilde{U}^{k-1}_y \Big(\tau^{-1} \int_\Omega \varphi_y\varphi_z \,\mathrm{d}x - (1-\theta) \int_\Omega \nabla\varphi_y \cdot \nabla\varphi_z \,\mathrm{d}x\Big)& \\
+ \sum_{y\in\mathscr{N}_h} \Big(- d_t u^k_{\mathrm{D},h}(y) \int_\Omega \varphi_y\varphi_z \,\mathrm{d}x - u^{k,\theta}_{\mathrm{D},h}(z) \int_\Omega \nabla\varphi_y \cdot \nabla\varphi_z \,\mathrm{d}x\Big)& \\
+ \sum_{y\in\mathscr{N}_h} \Big(f_h(t_{k,\theta}, y) \int_\Omega \varphi_y\varphi_z \,\mathrm{d}x + g_h(t_{k,\theta}, y) \int_{\Gamma_{\mathrm{N}}} \varphi_y\varphi_z \,\mathrm{d}x\Big)&
\end{aligned}$$

```
function [s,m,m_lumped,vol_T] = fe_matrices(c4n,n4e)
[nC,d] = size(c4n); nE = size(n4e,1);
m_loc = (ones(d+1,d+1)+eye(d+1))/((d+1)*(d+2));
ctr = 0; ctr_max = (d+1)^2*nE;
I = zeros(ctr_max,1); J = zeros(ctr_max,1);
X_s = zeros(ctr_max,1); X_m = zeros(ctr_max,1);
m_lumped_diag = zeros(nC,1); vol_T = zeros(nE,1);
for j = 1:nE
    X_T = [ones(1,d+1);c4n(n4e(j,:),:)'];
    grads_T = X_T\[zeros(1,d);eye(d)];
    vol_T(j) = det(X_T)/factorial(d);
    for m = 1:d+1
        for n = 1:d+1
            ctr = ctr+1; I(ctr) = n4e(j,m); J(ctr) = n4e(j,n);
            X_s(ctr) = vol_T(j)*grads_T(m,:)*grads_T(n,:)';
            X_m(ctr) = vol_T(j)*m_loc(m,n);
        end
        m_lumped_diag(n4e(j,m)) = m_lumped_diag(n4e(j,m))...
            +vol_T(j)/(d+1);
    end
end
s = sparse(I,J,X_s,nC,nC); m = sparse(I,J,X_m,nC,nC);
m_lumped = diag(m_lumped_diag);
```

```
function [m_bdy,m_lumped_bdy] = fe_matrices_bdy(c4n,bdy)
[nC,d] = size(c4n); n_bdy = size(bdy,1);
M_loc_bdy = (eye(d)+ones(d,d))/((d+1)*d);
ctr = 0; ctr_max = d^2*n_bdy;
I = zeros(ctr_max,1); J = zeros(ctr_max,1);
X_m_bdy = zeros(ctr_max,1);
m_lumped_bdy_diag = zeros(nC,1);
for j = 1:n_bdy
    if d == 1
        vol_S = 1;
    elseif d == 2
        vol_S = sqrt(sum((c4n(bdy(j,2),:)-c4n(bdy(j,1),:)).^2,2));
    elseif d == 3
        vol_S = sqrt(sum(cross(c4n(bdy(j,3),:)-c4n(bdy(j,1),:),...
            c4n(bdy(j,2),:)-c4n(bdy(j,1),:)).^2,2))/2;
    end
    for m = 1:d
        for n = 1:d
            ctr = ctr+1; I(ctr) = bdy(j,m); J(ctr) = bdy(j,n);
            X_m_bdy(ctr) = vol_S*M_loc_bdy(m,n);
         end
         m_lumped_bdy_diag(bdy(j,m)) = ...
             m_lumped_bdy_diag(bdy(j,m))+vol_S/d;
    end
end
m_bdy = sparse(I,J,X_m_bdy,nC,nC);
m_lumped_bdy = diag(m_lumped_bdy_diag);
```

Fig. 3.11 MATLAB routines that provide the $P1$ finite element stiffness and mass matrices according to Lemmas 3.10 and 3.12; the index "lumped" refers to reduced integration

for every $z \in \mathscr{K}_h$. The implementation thus requires computing L^2-inner products of the nodal basis functions. These can be replaced by simplified discrete versions based on numerical integration as introduced in Definition 3.13.

Lemma 3.12 (Mass matrices) *For $T \in \mathscr{T}_h$ such that $T = \operatorname{conv}\{z_0, z_1, \dots, z_d\}$, we have for $0 \le m, n \le d$ that*

$$\int_T \varphi_{z_m}\varphi_{z_n}\,\mathrm{d}x = \frac{|T|(1+\delta_{mn})}{(d+1)(d+2)}, \quad \int_T \mathscr{I}_h[\varphi_{z_m}\varphi_{z_n}]\,\mathrm{d}x = \frac{|T|\delta_{mn}}{d+1}.$$

For $S \in \mathscr{S}_h$ such that $S = \operatorname{conv}\{z_0, z_1, \dots, z_{d-1}\}$ we have for $0 \le m, n \le d-1$ that

$$\int_S \varphi_{z_m}\varphi_{z_n}\,\mathrm{d}x = \frac{|S|(1+\delta_{mn})}{d(d+1)}, \quad \int_S \mathscr{I}_h[\varphi_{z_m}\varphi_{z_n}]\,\mathrm{d}x = \frac{|S|\delta_{mn}}{d}.$$

Proof The identities follow from elementary calculations on a reference element and a transformation to T. □

Figure 3.10 displays a MATLAB implementation of the θ-midpoint scheme. The routines `fe_matrices_bdy.m` and `fe_matrices.m` displayed in Fig. 3.11 provide the stiffness and mass matrices. The parameter α in the code determines the time-step size via $\tau = h^\alpha/4$.

References

1. Alberty, J., Carstensen, C., Funken, S.A.: Remarks around 50 lines of Matlab: short finite element implementation. Numer. Algorithm. **20**(2–3), 117–137 (1999). http://dx.doi.org/10.1023/A:1019155918070
2. Braess, D.: Finite Elements, 3rd edn. Cambridge University Press, Cambridge (2007)
3. Brenner, S.C., Scott, L.R.: The Mathematical Theory of Finite Element Methods. Texts in Applied Mathematics, 3rd edn. Springer, New York (2008)
4. Chen, L.: iFEM: an integrated finite element methods package in MATLAB. Technical report, University of California Irvine. https://bitbucket.org/ifem/ifem/get/tip.zip (2009)
5. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. Classics in Applied Mathematics, vol. 40. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2002)
6. Dziuk, G.: Theorie und Numerik partieller Differentialgleichungen. Walter de Gruyter GmbH & Co. KG, Berlin (2010)
7. Gander, M.J., Wanner, G.: From Euler, Ritz, and Galerkin to modern computing. SIAM Rev. **54**(4), 627–666 (2012). http://dx.doi.org/10.1137/100804036
8. Hackbusch, W.: Theorie und Numerik elliptischer Differentialgleichungen, 2nd edn. Teubner Mathematical Textbooks, B. G. Teubner, Stuttgart (1996)
9. Larson, M.G., Bengzon, F.: The Finite Element Method: Theory, Implementation and Applications. Texts in Computational Science and Engineering, vol. 10. Springer, Heidelberg (2013)
10. Larsson, S., Thomée, V.: Partial Differential Equations with Numerical Methods. Texts in Applied Mathematics, vol. 45. Springer, Berlin (2003)
11. Rannacher, R.: Numerische Mathematik 2 (Numerik partieller Differentialgleichungen). http://numerik.iwr.uni-heidelberg.de/~lehre/notes/ (2008)
12. Thomée, V.: Galerkin Finite Element Methods for Parabolic Problems. Springer Series in Computational Mathematics, 2nd edn. Springer, Berlin (2006)

Chapter 4
Concepts for Discretized Problems

4.1 Convergence of Minimizers

We consider an abstract finite-dimensional minimization problem that seeks a minimizing function $u_h \in \mathscr{A}_h$ for a functional

$$I_h(u_h) = \int_\Omega W_h(\nabla u_h)\,\mathrm{d}x,$$

where the indices h in $\mathscr{A}_h$ and W_h refer to discretized versions of given counterparts in the infinite-dimensional variational problem for minimizing

$$I(u) = \int_\Omega W(\nabla u)\,\mathrm{d}x$$

in the set of functions $u \in \mathscr{A}$. We will often refer to the infinite-dimensional problem as the *continuous problem*, but this does not imply a continuity property of the functional or its integrand. The finite-dimensional problems will also be referred to as *discretized problems*. We recall that it is sufficient for the existence of discrete solutions to have coercivity and lower semicontinuity of I_h, while in the continuous situation, coercivity and the strictly stronger notion of weak lower semicontinuity of I are required. We discuss in this section the variational convergence of minimization problems and adopt concepts described in the textbook [5].

4.1.1 Failure of Convergence

A natural question to address is whether a family of discrete solutions $(u_h)_{h>0}$ converges to a minimizer $u \in \mathscr{A}$ for I with respect to some topology. Obviously, this requires the existence of a minimizer $u \in \mathscr{A}$ for I and convergence of the entire

S. Bartels, *Numerical Methods for Nonlinear Partial Differential Equations*,
Springer Series in Computational Mathematics 47,
DOI 10.1007/978-3-319-13797-1_4

sequence of approximations requires uniqueness of the continuous solution, or a certain selection principle contained in the discrete problems. Surprisingly, even if a solution exists for the continuous problem, if the discretization is conforming in the sense that $\mathscr{A}_h \subset \mathscr{A}$ and $W_h = W$, and if the family $(\mathscr{A}_h)_{h>0}$ is dense in $\mathscr{A}$, then convergence of discrete solutions may fail entirely.

Example 4.1 (*Lavrentiev phenomenon* [9]) Let $\mathscr{A}$ be the set of all functions $v \in W^{1,1}(0,1)$ satisfying $v(0) = 0$ and $v(1) = 1$ and consider

$$I(u) = \int_0^1 \big(x - u^3\big)^2 |u'|^6 \, \mathrm{d}x.$$

For $h > 0$ let $\mathscr{T}_h$ be a triangulation of $(0, 1)$, and define $\mathscr{A}_h = \mathscr{A} \cap \mathscr{S}^1(\mathscr{T}_h)$. Then the function $u(x) = x^{1/3}$ is a minimizer for I in $\mathscr{A}$, but for every $h > 0$, we have

$$0 = \min_{u \in \mathscr{A}} I(u) < \min_{u \in \mathscr{A} \cap W^{1,\infty}(0,1)} I(u) \le \min_{u_h \in \mathscr{A}_h} I(u_h).$$

In particular, the discrete minimal energies cannot converge to the right value. The reason for this discrepancy is the incompatibility of the growth of the integrand of I and the exponent of the employed Sobolev space in the definition of $\mathscr{A}$.

The example shows that even the seemingly simple notion of convergence

$$\min_{u_h \in \mathscr{A}_h} I_h(u_h) \to \inf_{u \in \mathscr{A}} I(u)$$

for $h \to 0$ requires stronger arguments than just the density of the approximation spaces. Once convergence is understood, a natural question to investigate is whether a rate of convergence can be proved, i.e., whether there exists $\alpha > 0$ with

$$\Big| \min_{u_h \in \mathscr{A}_h} I_h(u_h) - \inf_{u \in \mathscr{A}} I(u) \Big| \le c h^{\alpha}.$$

Even if this is the case, it is not guaranteed that discrete solutions $u_h \in \mathscr{A}_h$ converge to a minimizer $u \in \mathscr{A}$ of I.

Example 4.2 (*Lack of weak lower semicontinuity*) Set $\mathscr{A} = W^{1,4}(0,1)$ and let

$$I(u) = \int_0^1 \big(|u'|^2 - 1\big)^2 + u^4 \, \mathrm{d}x.$$

For $h > 0$ let $\mathscr{T}_h$ be a triangulation of $(0, 1)$ of maximal mesh-size h and define $\mathscr{A}_h = \mathscr{A} \cap \mathscr{S}^1(\mathscr{T}_h)$. Then $\inf_{u \in \mathscr{A}} I(u) = 0$ and

$$\Big| \min_{u_h \in \mathscr{A}_h} I(u_h) - \inf_{u \in \mathscr{A}} I(u) \Big| \le c h^4,$$

and any weakly convergent sequence of discrete minimizers $(u_h)_{h>0}$ satisfies $u_h \rightharpoonup 0$ in $W^{1,4}(\Omega)$ as $h \to 0$. Due to the nonconvexity of the integrand, we have that $u = 0$ is not a minimizer for I, i.e., $0 < 1 = I(0)$.

4.1.2 *Γ-Convergence of Discretizations*

The concept of Γ-convergence provides a concise framework to analyze convergence of a sequence of energy functionals and its minimizers. In an abstract form we consider a sequence of discrete minimization problems:

$$\text{Minimize } I_h(u_h) \text{ in the set of functions } u_h \in X_h.$$

Here, every space X_h is assumed to be a subspace of a Banach space X and I_h is allowed to attain the value $+\infty$, so that constraints contained in $\mathscr{A}_h \subset X_h$ can be incorporated in I_h. We formally extend the discrete problems to X by setting

$$I_h(u) = \begin{cases} I_h(u) & \text{if } u \in X_h, \\ +\infty & \text{if } u \notin X_h. \end{cases}$$

In the following, $h > 0$ stands for a sequence of positive real numbers that accumulate at zero.

Definition 4.1 Let X be a Banach space, $I : X \to \mathbb{R} \cup \{+\infty\}$, and let $(I_h)_{h>0}$ be a sequence of functionals $I_h : X \to \mathbb{R} \cup \{+\infty\}$. We say that *the sequence* $(I_h)_{h>0}$ *Γ-converges to I* as $h \to 0$, denoted by $I_h \to^{\Gamma} I$, with respect to a given topology ω on X if the following conditions hold:

(a) For every sequence $(u_h)_{h>0} \subset X$ with $u_h \to^{\omega} u$ for some $u \in X$, we have that $\liminf_{h\to 0} I_h(u_h) \geq I(u)$.
(b) For every $u \in X$ there exists a sequence $(u_h)_{h>0} \subset X$ with $u_h \to^{\omega} u$ and $I_h(u_h) \to I(u)$ as $h \to 0$.

Remark 4.1 The first condition is called liminf-inequality and implies that I is a lower bound for the sequence $(I_h)_{h>0}$ in the limit $h \to 0$. The second condition guarantees that the lower bound is attained, and the involved sequence is called a recovery sequence.

Unless otherwise stated, we consider the weak topology ω on X. For conforming discretizations, i.e., if $I_h(u_h) = I(u_h)$ for all $u_h \in X_h$, of well-posed minimization problems, a Γ-convergence result can be proved under moderate conditions.

Theorem 4.1 (Conforming discretizations) *Assume that $I_h(u_h) = I(u_h)$ for $u_h \in X_h$ and $h > 0$ and that the spaces $(X_h)_{h>0}$ are dense in X with respect to the strong topology of X. If I is weakly lower semicontinuous and strongly continuous, then we have $I_h \to^{\Gamma} I$ as $h \to 0$ with respect to weak convergence in X.*

Proof Let $(u_h)_{h>0} \subset X$ and $u \in X$ be such that $u_h \rightharpoonup u$ as $h \to 0$. To prove the liminf-inequality, we note that $I_h(u_h) \geq I(u_h)$ and thus the weak lower semicontinuity of I implies $\liminf_{h\to 0} I_h(u_h) \geq \liminf_{h\to 0} I(u_h) \geq I(u)$. To prove that $I(u)$ is attained for every $u \in X$, let $(u_h)_{h>0}$ be a sequence with $u_h \in X_h$ for every $h > 0$ and $u_h \to u$ in X. The strong continuity of I and $I_h(u_h) = I(u_h)$ imply that $I(u) = \lim_{h\to 0} I_h(u_h)$. □

The definition of Γ-convergence has remarkable consequences.

Proposition 4.1 (Γ-Convergence)

(i) *If $I_h \to^{\Gamma} I$ as $h \to 0$, then I is weakly lower semicontinuous on X.*
(ii) *If $I_h \to^{\Gamma} I$ as $h \to 0$ and for every $h > 0$ there exists $u_h \in X$ such that $I_h(u_h) \leq \inf_{v_h\in X} I_h(v_h) + \varepsilon_h$ with $\varepsilon_h \to 0$ as $h \to 0$ and $u_h \to^{\omega} u$ for some $u \in X$, then $I_h(u_h) \to I(u)$ and u is a minimizer for I.*
(iii) *If $I_h \to^{\Gamma} I$ and G is ω-continuous on X, then $I_h + G \to^{\Gamma} I + G$.*

Proof (i) Let $(u_j)_{j\in\mathbb{N}} \subset X$ be a sequence with $u_j \to^{\omega} u$ in X as $j \to \infty$. For every $j \in \mathbb{N}$ there exists a sequence $(u_j^h)_{h>0}$ such that $u_j^h \to^{\omega} u_j$ as $h \to 0$ and $I_h(u_j^h) \to I(u_j)$. For every $j \in \mathbb{N}$ we may thus choose $h_j > 0$, such that $|I(u_j) - I_{h_j}(u_j^{h_j})| \leq 1/j$ and $u_j^{h_j} \to^{\omega} u$ as $j \to \infty$. It follows that

$$I(u) \leq \liminf_{j\to\infty} I_{h_j}\big(u_j^{h_j}\big) = \liminf_{j\to\infty} I(u_j) - I(u_j) + I_{h_j}\big(u_j^{h_j}\big) = \liminf_{j\to\infty} I(u_j).$$

This proves the first statement.
(ii) If $u_h \to^{\omega} u$, then by condition (a) we have $I(u) \leq \liminf_{h\to 0} I_h(u_h)$. Moreover, due to (b) for every $v \in X$, there exists $(v_h)_{h>0} \subset X$ with $v_h \to^{\omega} v$ and $I_h(v_h) \to I(v)$ as $h \to 0$. Therefore, $I(u_h) \leq I(v_h) + \varepsilon_h$ and

$$I(u) \leq \liminf_{h\to 0} I_h(u_h) \leq \lim_{h\to 0} \big(I_h(v_h) + \varepsilon_h\big) = I(v),$$

i.e., u is a minimizer for I.
(iii) If G is ω-continuous, then $G(u_h) \to G(u)$ whenever $u_h \to^{\omega} u$ in X and the Γ-convergence of $I_h + G$ to $I + G$ follows directly from $I_h \to^{\Gamma} I$. □

4.1.3 Examples of Γ-Convergent Discretizations

We discuss some examples of Γ-convergence. As above, we always extend a functional I_h defined on a subspace $X_h \subset X$ by the value $+\infty$ to the whole space X.

Example 4.3 (*Poisson problem*) Let $X = H^1_{\mathrm{D}}(\Omega)$ and $X_h = \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ for a regular family of triangulations $(\mathscr{T}_h)_{h>0}$ of Ω. For $f \in L^2(\Omega)$ and $g \in L^2(\Gamma_{\mathrm{N}})$, let

$$I(u) = \frac{1}{2}\int_{\Omega} |\nabla u|^2 \,\mathrm{d}x - \int_{\Omega} fu \,\mathrm{d}x - \int_{\Gamma_{\mathrm{N}}} gu \,\mathrm{d}s$$

and let $I_h : H^1_{\mathrm{D}}(\Omega) \to \mathbb{R} \cup \{+\infty\}$ coincide with I on $\mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$. Since the Dirichlet energy is weakly lower semicontinuous and strongly continuous, the linear lower-order terms are weakly continuous on $H^1_{\mathrm{D}}(\Omega)$, and since the finite element spaces are dense in $H^1_{\mathrm{D}}(\Omega)$, we verify that $I_h \to^{\Gamma} I$ as $h \to 0$. Nonhomogeneous Dirichlet conditions can be included by considering the decomposition $u = \widetilde{u} + \widetilde{u}_{\mathrm{D}}$ with $\widetilde{u} \in H^1_{\mathrm{D}}(\Omega)$. For minimizers $u \in H^2(\Omega) \cap H^1_{\mathrm{D}}(\Omega)$ of I and $u_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ of I_h, we have

$$\big|I(u) - I_h(u_h)\big| \le ch.$$

A constant sequence of functionals can have a different Γ-limit.

Example 4.4 (*Relaxation*) For the sequence of functionals defined through $X = W^{1,4}(0,1)$,

$$I(u) = \int_0^1 \big(|u'|^2 - 1\big)^2 + u^4 \,\mathrm{d}x,$$

subspaces $X_h = \mathscr{S}^1(\mathscr{T}_h)$, and $I_h = I$ on X_h, we have that $I_h \to^{\Gamma} I^{**}$ in $W^{1,4}(0,1)$ with the convexified functional

$$I^{**}(u) = \int_0^1 \big(|u'|^2 - 1\big)_+^2 + u^4 \,\mathrm{d}x,$$

where $s_+ = \max\{s, 0\}$ for $s \in \mathbb{R}$. Since the integrand of I^{**} is convex, the functional is weakly lower semicontinuous. Using that $I_h(u_h) = I(u_h) \ge I^{**}(u_h)$ for all $h > 0$, we deduce that $\liminf_{h\to 0} I_h(u_h) \ge I^{**}(u)$ whenever $u_h \rightharpoonup u$ in $W^{1,4}(0,1)$. To prove that the lower bound is attained, we first consider the case that $u \in W^{1,4}(\Omega)$ is piecewise affine, i.e., $u = u_H \in \mathscr{S}^1(\mathscr{T}_H)$ for some $H > 0$. For $0 < h < H$ we then construct a function u_h that nearly coincides with u_H on elements $T_H \in \mathscr{T}_H$ for which $|u'_H|_{T_H}| \ge 1$. For elements with $|u'_H|_{T_H}| \le 1$ we use gradients $u'_h \in \{\pm 1\}$ on T_H in such a way that u_h and u_H nearly coincide at the endpoints of T_H and differ by at most h in the interior. Then $I(u_h) \approx I^{**}(u_H)$ and $I(u_h) \to I^{**}(u_H)$ as $h \to 0$. The construction is depicted in Fig. 4.1. The assertion for general $u \in W^{1,4}(\Omega)$ follows from an approximation result and the strong continuity of I.

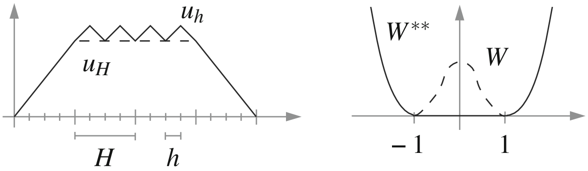


Fig. 4.1 Construction of an oscillating function u_h (*solid line*) with $|u'_h| \ge 1$ that approximates u_H (*dashed line*) such that $I(u_h) \approx I^{**}(u_H)$ (*left*) in Example 4.4; the integrand W^{**} (*solid line*) of I^{**} is the convex hull of the integrand W (*dashed line*) of I (*right*)

A typical application of conforming discretizations of well-posed minimization problems occurs in simulating hyperelastic materials.

Example 4.5 (*Hyperelasticity*) Let $\mathscr{A} = \{y \in W^{1,p}(\Omega;\mathbb{R}^d) : y|_{\Gamma_{\rm D}} = \widetilde{y}_{\rm D}|_{\Gamma_{\rm D}}\}$ for $1 \le p < \infty$ and $\widetilde{y}_{\rm D} \in W^{1,p}(\Omega;\mathbb{R}^d)$. Assume that $W : \mathbb{R}^{d\times d} \to \mathbb{R}$ is continuous and quasiconvex with

$$-c_1 + c_2|F|^p \le W(F) \le c_1 + c_2|F|^p.$$

Then for $f \in L^{p'}(\Omega;\mathbb{R}^d)$ and $g \in L^{p'}(\Gamma_{\rm N};\mathbb{R}^d)$, the functional

$$I(y) = \int\limits_{\Omega} W(\nabla y)\,{\rm d}x - \int\limits_{\Omega} f\cdot y\,{\rm d}x - \int\limits_{\Gamma_{\rm N}} g\cdot y\,{\rm d}s$$

is weakly lower semicontinuous and coercive on $W^{1,p}(\Omega;\mathbb{R}^d)$. Moreover, if the sequence $(y_j)_{j\in\mathbb{N}} \subset W^{1,p}(\Omega;\mathbb{R}^d)$ converges strongly to $y \in W^{1,p}(\Omega;\mathbb{R}^d)$ then we have $\nabla y_{j_k}(x) \to \nabla y(x)$ for almost every $x \in \Omega$ for a subsequence $(y_{j_k})_{k\in\mathbb{N}}$, and the generalized dominated convergence theorem implies

$$\int\limits_{\Omega} W(\nabla y_{j_k})\,{\rm d}x \to \int\limits_{\Omega} W(\nabla y)\,{\rm d}x,$$

i.e., up to subsequences I is strongly continuous and this is sufficient to establish Γ-convergence. For piecewise affine boundary data $y_{\rm D}$, we have that $\mathscr{A}_h = \mathscr{A} \cap \mathscr{S}^1(\mathscr{T}_h)^d$ is nonempty and the density of finite element spaces implies $I_h \to^{\Gamma} I$ for conforming discretizations. More generally, it suffices to consider convergent approximations $\widetilde{y}_{{\rm D},h}$ of $\widetilde{y}_{\rm D}$.

The abstract convergence theory allows us to include nonlinear constraints.

Example 4.6 (*Harmonic maps*) Assume that $u_{\rm D} \in C(\Gamma_{\rm D};\mathbb{R}^m)$ is such that

$$\mathscr{A} = \{u \in H^1(\Omega;\mathbb{R}^m) : u|_{\Gamma_{\rm D}} = u_{\rm D},\ |u(x)| = 1 \text{ f.a.e. } x \in \Omega\}$$

is nonempty and for a triangulation $\mathscr{T}_h$ of Ω with nodes $\mathscr{N}_h$, set

$$\mathscr{A}_h = \{u_h \in \mathscr{S}^1(\mathscr{T}_h)^m : u(z) = u_{\rm D}(z) \text{ f.a. } z \in \mathscr{N}_h \cap \Gamma_{\rm D},\ |u_h(z)| = 1 \text{ f.a. } z \in \mathscr{N}_h\},$$

i.e., $\mathscr{A}_h \not\subset \mathscr{A}$. We then consider the minimization of the Dirichlet energy I on $\mathscr{A}_h$ and $\mathscr{A}$, respectively, which defines minimization problems with functionals I_h and I on $H^1(\Omega;\mathbb{R}^m)$, respectively. To show that $I_h \to^{\Gamma} I$ in $H^1(\Omega;\mathbb{R}^m)$ we note that the liminf-inequality follows from the weak lower semicontinuity of I, together with the fact that if $u_h \rightharpoonup u$ in $W^{1,2}(\Omega;\mathbb{R}^m)$ with $u_h \in \mathscr{A}_h$ for every $h > 0$, then $u \in \mathscr{A}$. The latter implication follows from a nodal interpolation result, together with elementwise inverse estimates, i.e.,

$$\||u_h|^2 - 1\| = \||u_h|^2 - \mathscr{I}_h|u_h|^2\| \le ch\|u_h\|\|\nabla u_h\|.$$

Therefore, $|u_{h'}(x)| \to 1$ for almost every $x \in \Omega$ and a subsequence $h' > 0$ so that $|u(x)| = 1$ for almost every $x \in \Omega$. We assume that u_{D} is sufficiently regular, so that a similar argument shows $u|_{\Gamma_{\mathrm{D}}} = u_{\mathrm{D}}$. To prove the attainment of I, we note that due to the density of smooth unit-length vector fields in $\mathscr{A}$, we may assume $u \in \mathscr{A} \cap H^2(\Omega;\mathbb{R}^m)$ and define $u_h = \mathscr{I}_h u \in \mathscr{A}_h$. Then $u_h \to u$ in $H^1(\Omega;\mathbb{R}^m)$ and $I_h(u_h) \to I(u)$ as $h \to 0$.

Remark 4.2 In general, smooth constrained vector fields are not dense in sets of weakly differentiable constrained vector fields, cf., e.g., [18].

For practical purposes it is often desirable to modify a given functional.

Example 4.7 (*Total variation minimization*) For $X = W^{1,1}(\Omega)$ we consider

$$I(u) = \int_\Omega |\nabla u| \,\mathrm{d}x;$$

and given a family of triangulations $(\mathscr{T}_h)_{h>0}$ of Ω and $u_h \in \mathscr{S}^1(\mathscr{T}_h)$, we define for $\beta > 0$ the regularized functionals

$$I_h(u_h) = \int_\Omega (h^\beta + |\nabla u_h|^2)^{1/2} \,\mathrm{d}x.$$

If $u_h \rightharpoonup u$ in $W^{1,1}(\Omega)$, then the liminf-inequality follows from the weak lower semicontinuity of I on $W^{1,1}(\Omega)$ and the fact that $I_h(u_h) \geq I(u_h)$ for every $h > 0$. To verify that $I(u)$ is attained for every $u \in W^{1,1}(\Omega)$ in the limit $h \to 0$, we note that the density of finite element spaces in $W^{1,1}(\Omega)$ allows us to consider a sequence $(u_h)_{h>0} \subset W^{1,1}(\Omega)$ with $u_h \in \mathscr{S}^1(\mathscr{T}_h)$ for every $h > 0$ and $u_h \to u \in W^{1,1}(\Omega)$ as $h \to 0$. The estimate $(a^2+b^2)^{1/2} \leq |a| + |b|$ implies that

$$(h^\beta + |\nabla u_h|^2)^{1/2} - |\nabla u| \leq h^{\beta/2} + |\nabla u_h| - |\nabla u|,$$

and for a subsequence we have $((h')^\alpha + |\nabla u_{h'}|^2)^{1/2} \to |\nabla u|$ almost everywhere in Ω. The generalized dominated convergence theorem implies that $I_{h'}(u_{h'}) \to I(u)$ as $h' \to 0$. With Proposition 4.1, this also implies the Γ-convergence of discretizations of

$$I(u) = \int_\Omega |\nabla u| \,\mathrm{d}x + \frac{\alpha}{2}\|u-g\|^2$$

for $g \in L^2(\Omega)$. Due to the lack of reflexivity of $W^{1,1}(\Omega)$ this is not sufficient to deduce the existence of minimizers for I, i.e., we cannot deduce the existence of weak limits of (subsequences) of a bounded sequence. For this, the larger space $BV(\Omega) \cap L^2(\Omega)$ has to be considered. A corresponding Γ-convergence result follows analogously with the density of $W^{1,1}(\Omega)$ in $BV(\Omega)$ with respect to an appropriate notion of convergence.

4.1.4 Error Control for Strongly Convex Problems

For Banach spaces X and Y, a bounded linear operator $\Lambda : X \to Y$, and convex, lower-semicontinuous, proper functionals $F : X \to \mathbb{R} \cup \{+\infty\}$ and $G : Y \to \mathbb{R} \cup \{+\infty\}$, we consider the problem of finding $u \in X$ with

$$I(u) = \inf_{v \in X} I(v), \quad I(v) = F(v) + G(\Lambda v).$$

The *Fenchel conjugates* $F^* : X' \to \mathbb{R} \cup \{+\infty\}$ and $G^* : Y' \to \mathbb{R} \cup \{+\infty\}$ are the convex, lower-semicontinuous, proper functionals defined by

$$F^*(w) = \sup_{v \in X} \langle w, v \rangle - F(v), \quad G^*(q) = \sup_{p \in Y} \langle q, p \rangle - G(p)$$

for $w \in X'$ and $q \in Y'$, respectively. We assume that Y is reflexive, so that $G = G^{**}$. Then, the property of the formal adjoint operator $\Lambda' : Y' \to X'$, that $\langle \Lambda v, q \rangle = \langle v, \Lambda' q \rangle$, and the general relation $\inf_v \sup_q H(v, q) \geq \sup_q \inf_v H(v, q)$ for an arbitrary function $H : X \times Y' \to \mathbb{R} \cup \{+\infty\}$ yield

$$\begin{aligned}\inf_v I(v) &= \inf_v F(v) + G^{**}(\Lambda v) = \inf_v \sup_q F(v) + \langle v, \Lambda' q \rangle - G^*(q) \\ &\geq \sup_q \inf_v F(v) + \langle v, \Lambda' q \rangle - G^*(q) = \sup_q \inf_v F(v) - \langle v, -\Lambda' q \rangle - G^*(q) \\ &= \sup_q \Big(- \sup_v \langle v, -\Lambda' q \rangle - F(v) - G^*(q) \Big) = \sup_q -F^*(-\Lambda' q) - G^*(q).\end{aligned}$$

This motivates considering the dual problem which consists in finding $p \in Y'$ with

$$D(p) = \sup_{q \in Y'} D(q), \quad D(q) = -F^*(-\Lambda' q) - G^*(q).$$

We assume that F or G is *strongly convex*, so that there exist $\alpha_F, \alpha_G \geq 0$ with $\max\{\alpha_F, \alpha_G\} > 0$, so that for all $q_1, q_2 \in Y$ and $v_1, v_2 \in X$, we have

$$\begin{aligned}G\big((q_1 + q_2)/2\big) + \alpha_G \|q_2 - q_1\|_Y^2 &\leq \frac{1}{2}\big(G(q_1) + G(q_2)\big), \\ F\big((v_1 + v_2)/2\big) + \alpha_F \|v_2 - v_1\|_X^2 &\leq \frac{1}{2}\big(F(v_1) + F(v_2)\big).\end{aligned}$$

By convexity, the estimates hold with $\alpha_G = \alpha_F = 0$. The primal and dual optimization problems are related by the weak *complementarity principle*

$$I(u) = \inf_{v \in X} I(v) \geq \sup_{q \in Y^*} D(q) = D(p).$$

We say that *strong duality* applies if equality holds. Our final ingredient for the error estimate is a characterization of the optimality of the solution of the primal problem.

For some $\alpha_I \geq 0$ and all $w \in \partial I(u)$, we have that

$$\langle w, v-u\rangle + \alpha_I \|v-u\|_X^2 \leq I(v) - I(u)$$

and u is optimal if and only if $0 \in \partial I(u)$. We assume in the following that $\alpha_F > 0$ or $\alpha_I > 0$, so that I has a unique minimizer $u \in X$.

Theorem 4.2 (Error control [16]) *Assume that* $\max\{\alpha_F, \alpha_G, \alpha_I\} > 0$ *and let* $u \in X$ *be the unique minimizer for* I.

(i) *For a minimizer* $u_h \in X_h$ *for* I *restricted to a subspace* $X_h \subset X$, *we have the a priori error estimate*

$$\alpha_G \|\Lambda(u-u_h)\|_Q^2 + (\alpha_F + \alpha_I/4)\|u-u_h\|_X^2 \leq \inf_{w_h \in X_h} \frac{1}{2}\big(I(w_h) - I(u)\big).$$

(ii) *For an arbitrary approximation* $\widetilde{u}_h \in X$ *of* u, *we have the a posteriori error estimate*

$$\alpha_G \|\Lambda(u-\widetilde{u}_h)\|_Q^2 + (\alpha_F + \alpha_I/4)\|u-\widetilde{u}_h\|_X^2 \leq \inf_{q \in Y'} \frac{1}{2}\big(I(\widetilde{u}_h) - D(q)\big).$$

Proof The convexity estimates imply that

$$\alpha_G \|\Lambda(u-v)\|_Q^2 + \alpha_F \|u-v\|_X^2 \leq \frac{1}{2}\big(I(v) + I(u)\big) - I\big((v+u)/2\big).$$

The optimality of u shows that we have

$$I(u) + \alpha_I \|u-(u+v)/2\|_X^2 \leq I\big((u+v)/2\big).$$

It follows that

$$\alpha_G \|\Lambda(u-v)\|_Q^2 + \alpha_F \|u-v\|_X^2 \leq \frac{1}{2}\big(I(v) - I(u)\big) - \alpha_I \|((u-v)/2\|_X^2.$$

If $u_h \in X_h$ is minimal in X_h, then the identity $I(u_h) = \inf_{w_h \in X_h} I(w_h)$ implies the a priori estimate. The weak complementarity principle $I(u) \geq D(q)$ yields the a posteriori estimate. □

Remarks 4.3 (i) If *strong duality* holds, i.e., if $I(u) = D(p)$, then the estimate of the theorem is sharp in the sense that the right-hand side vanishes if $v = u$ and q solves the dual problem.
(ii) Sufficient conditions for strong duality are provided by von Neumann's minimax theorem, e.g., that F and G^* are convex, lower semicontinuous, and coercive.

Example 4.8 For the Poisson problem $-\Delta u = f$ in Ω, $u|_{\partial\Omega} = 0$, we have $X = H_0^1(\Omega)$, $Y = L^2(\Omega;\mathbb{R}^d)$, $\Lambda = \nabla$, $G(\Lambda v) = (1/2)\int_\Omega |\nabla v|^2 \,\mathrm{d}x$, and

$F(v) = -\int_\Omega f v \,\mathrm{d}x$. It follows that $F^*(w) = I_{\{-f\}}(w)$, $G^*(q) = (1/2)\int_\Omega |q|^2 \,\mathrm{d}x$,

$$\Lambda' = -\operatorname{div} : L^2(\Omega; \mathbb{R}^d) \to H_0^1(\Omega)^*.$$

We thus have

$$\frac{1}{2}\big((q_1+q_2)/2\big)^2 - \frac{1}{4}(q_1^2+q_2^2) = \frac{1}{8}(q_1^2+2q_1q_2+q_2^2-2q_1^2-2q_2^2) = -\frac{1}{8}(q_1-q_2)^2,$$

so that $\alpha_G = 1/8$ and

$$\frac{1}{2}q_1^2 - \frac{1}{2}q_2^2 - q_1(q_1-q_2) = -\frac{1}{2}q_1^2 - \frac{1}{2}q_2^2 + q_1q_2 = -\frac{1}{2}(q_1-q_2)^2,$$

i.e., $\alpha_I = 1/2$. Moreover, we have $\alpha_F = 0$.

(i) Incorporating the definition of the exact weak solution, the abstract a priori estimate of Theorem 4.2 provides the bound

$$\begin{aligned}\frac{1}{2}\|\nabla(u-u_h)\|^2 &\le \frac{1}{2}\int_\Omega |\nabla w_h|^2 - \int_\Omega f w_h \,\mathrm{d}x - \frac{1}{2}\int_\Omega |\nabla u|^2 + \int_\Omega f u \,\mathrm{d}x \\ &= \frac{1}{2}\|\nabla(u-w_h)\|^2 + \int_\Omega \nabla u \cdot \nabla(u-w_h)\,\mathrm{d}x + \int_\Omega f(u-w_h)\,\mathrm{d}x \\ &= \frac{1}{2}\|\nabla(u-w_h)\|^2,\end{aligned}$$

which implies the best-approximation property

$$\|\nabla(u-u_h)\| \le \inf_{w_h \in X_h} \|\nabla(u-w_h)\|.$$

(ii) Letting $\eta^2(v,q)$ denote the right-hand side of the a posteriori error estimate of Theorem 4.2, we have

$$\begin{aligned}2\eta^2(v,q) &= -\int_\Omega f v \,\mathrm{d}x + I_{\{-f\}}(\operatorname{div} q) + \frac{1}{2}\int_\Omega |\nabla v|^2 \,\mathrm{d}x + \frac{1}{2}\int_\Omega |q|^2 \,\mathrm{d}x \\ &= \int_\Omega (\operatorname{div} q) v \,\mathrm{d}x + \frac{1}{2}\|\nabla v\|^2 + \frac{1}{2}\|q\|^2 = \frac{1}{2}\|\nabla v - q\|^2,\end{aligned}$$

provided that $-\operatorname{div} q = f$. The theorem thus implies

$$\|\nabla(u-v)\| \le \inf_{-\operatorname{div} q = f} \|\nabla v - q\|.$$

4.2 Approximation of Equilibrium Points

The Euler–Lagrange equations related to a minimization problem typically seek a function $u \in X$ such that

$$F(u)[v] = \ell(v)$$

for all $v \in X$ with a possibly nonlinear operator $F : X \to X'$ and a linear functional $\ell \in X'$. Various other mathematical problems that may not be related to a minimization problem can also be formulated in this abstract form. A natural discretization employs subspaces $X_h \subset X$ and seeks $u_h \in X_h$ with

$$F_h(u_h)[v_h] = \ell_h(v_h)$$

for all $v_h \in X_h$. Here, $F_h : X_h \to X_h'$ and $\ell_h \in X_h'$ are approximations of F and ℓ that result from a discretization, e.g., via numerical integration. The important question to address is whether numerical solutions $(u_h)_{h>0}$ for a sequence of finite-dimensional subspaces X_h converge in an appropriate sense to a solution of the infinite-dimensional problem. We assume that the finite-dimensional space X_h is equipped with the norm of X. The corresponding dual spaces X_h' and X' are related by the inclusion $X'|_{X_h} \subset X_h'$. Topics related to the contents of this section can be found in the textbooks [3, 11].

4.2.1 *Failure of Convergence*

The following examples show that unjustified regularity assumptions can lead to the failure of convergence to the correct object. The following examples are taken from [6].

Example 4.9 (*Maxwell's equations*) For $\Omega \subset \mathbb{R}^2$ set $X = H_0(\mathrm{curl}; \Omega) \cap H(\mathrm{div}; \Omega)$, where

$$H_0(\mathrm{curl}; \Omega) = \{v \in L^2(\Omega; \mathbb{R}^2) : \mathrm{curl}\ v \in L^2(\Omega; \mathbb{R}^2),\ v \cdot t = 0 \text{ on } \partial\Omega\}$$

with $\mathrm{curl}\ v = \partial_1 v_2 - \partial_2 v_1$ for $v = (v_1, v_2)$ and $t : \partial\Omega \to \mathbb{R}^2$ a unit tangent. For $f \in L^2(\Omega; \mathbb{R}^2)$, consider the problem of finding $u \in X$ such that

$$(\mathrm{curl}\ u, \mathrm{curl}\ v) + (\mathrm{div}\ u, \mathrm{div}\ v) = (f, v)$$

for all $v \in X$. The existence and uniqueness of a solution follows from the Lax–Milgram lemma. A discretization of this problem is obtained by choosing $X_h = \mathscr{S}^1(\mathscr{T}_h)^2 \cap X$ and computing $u_h \in X_h$ such that

$$(\mathrm{curl}\ u_h, \mathrm{curl}\ v_h) + (\mathrm{div}\ u_h, \mathrm{div}\ v_h) = (f, v_h)$$

for all $v_h \in X_h$. This defines a convergent numerical scheme if Ω is convex. If Ω is nonconvex, then $H^1(\Omega; \mathbb{R}^2) \cap X$ is a closed proper subspace of X, cf. [8] for details, and convergence $u_h \to u$ as $h \to 0$ fails in general.

A similar effect occurs for higher-order problems.

Example 4.10 (*Biharmonic equation*) The biharmonic equation

$$\Delta^2 u = f \text{ in } \Omega, \quad u = \Delta u = 0 \quad \text{on } \partial\Omega$$

formally corresponds to the weak formulation that seeks $u \in H^2(\Omega) \cap H_0^1(\Omega)$ with

$$\int_\Omega D^2 u : D^2 v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x$$

for all $v \in H^2(\Omega) \cap H_0^1(\Omega)$ We denote the unique weak solution of the variational formulation by $u = (\Delta^2)^{-1} f$. A natural discretization of the problem is based on an operator splitting which is obtained by introducing $z = -\Delta u$ and solving the Poisson problems

$$\begin{aligned} -\Delta z &= f \text{ in } \Omega, \quad z = 0 \text{ on } \partial\Omega, \\ -\Delta u &= z \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega. \end{aligned}$$

We have $z = (-\Delta)^{-1} f$ and $u = (-\Delta)^{-1} z = (-\Delta)^{-2} f$. Unless Ω is convex so that $\Delta u \in H_0^1(\Omega)$ we do not have $(\Delta^2)^{-1} f = (-\Delta)^{-2} f$, and convergence of related numerical methods will fail in general.

Failure of convergence may also be related to the lack of uniqueness of a solution as in the case of degenerately monotone problems.

Example 4.11 (*Degenerate monotonicity*) For $\sigma(F) = DW^{**}(F)$ for $F \in \mathbb{R}^d$ and $W^{**}(F) = (|F|^2 - 1)_+^2$, there are infinitely many functions $u \in W_0^{1,4}(\Omega)$ satisfying $F(u)[v] = \int_\Omega \sigma(\nabla u) \cdot \nabla v \, \mathrm{d}x = 0$ for all $v \in W_0^{1,4}(\Omega)$.

4.2.2 Abstract Error Estimates

We sketch below the classical concept that consistency and stability imply the convergence of numerical approximations, provided that appropriate regularity results are available. Dual to this is an approach that leads to computable upper bounds for the approximation error and which avoids regularity assumptions entirely.

Theorem 4.3 (Abstract a priori error estimate) *Let $u \in X$ satisfy $F(u) = \ell$ and assume that for an interpolant $i_h u \in X_h$ and a consistency functional $\mathscr{C}_h(u) \in X_h'$, we have*

$$F_h(i_h u)[v_h] - \ell_h(v_h) = \mathscr{C}_h(u; v_h)$$

for all $v_h \in X_h$. Assume that we have discrete stability in the sense that for all $z_h \in X_h$ and $b_h \in X_h'$, the implication

$$\forall v_h \in X_h \;\; F_h(z_h)[v_h] = b_h(v_h) \implies \|z_h\|_X \leq c_{S,h} \|b_h\|_{X_h'}$$

holds. Then, if $F_h : X_h \to X_h'$ is linear, there exists a unique solution $u_h \in X_h$ with

$$\|u_h - i_h u\|_X \le c_{S,h} \|\mathscr{C}_h(u)\|_{X_h'}.$$

Proof Discrete stability implies that $F_h : X_h \to X_h'$ is a bijection and hence there exists a unique $u_h \in X_h$ with $F_h(u_h) = 0$. Since $F_h(i_h u - u_h) = F_h(i_h u) - F_h(u_h) = F_h(i_h u) - \ell_h = \mathscr{C}_h(u)$ we deduce the estimate. □

Remark 4.4 We say that a discretization is consistent of order $\beta \ge 0$, given the regularity $u \in Z \subset X$ if $\|\mathscr{C}_h(u)\|_{X_h'} \le ch^\beta$. This implies convergence of approximations with rate β.

A similar abstract concept leads to a posteriori error estimates for many linear problems.

Theorem 4.4 (Abstract a posteriori error estimate) *Let $u_h \in X_h$ and define the residual $\mathscr{R}_h(u_h) \in X'$ through*

$$\mathscr{R}_h(u_h; v) = F(u_h)[v] - \ell(v)$$

for all $v \in X$. Assume that we have the continuous stability result that for all $z \in X$ and $b \in X'$, the implication

$$\forall\, v \in X \;\; F(z)[v] = b(v) \implies \|z\|_X \le c_S \|b\|_{X'}$$

holds. If $u \in X$ satisfies $F(u) = \ell$ and if F is linear, then u is unique with

$$\|u - u_h\|_X \le c_S \|\mathscr{R}_h(u_h)\|_{X'}.$$

Proof The difference $u - u_h$ satisfies $F(u - u_h)[v] = \mathscr{R}_h(u_h; v)$ for all $v \in X$, and the stability result implies the error estimate and the uniqueness property. □

Example 4.12 (*Poisson problem*) Let $u \in H^1_{\mathrm{D}}(\Omega)$ be the weak solution of $-\Delta u = f$ in Ω, $u|_{\Gamma_{\mathrm{D}}} = 0$, and $\partial_\nu u|_{\Gamma_{\mathrm{N}}} = g$, i.e., we have $F(u) = \ell$ with

$$F(u)[v] = \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x, \quad \ell(v) = \int_\Omega f v \,\mathrm{d}x + \int_{\Gamma_{\mathrm{N}}} g v \,\mathrm{d}s.$$

The lowest-order finite element method seeks $u_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ with $F(u_h)[v_h] = \ell(v_h)$ for all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$.

(i) Inserting an interpolant $i_h u \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ in the discrete formulation leads to

$$\mathscr{C}_h(u; v_h) = F(i_h u)[v_h] - \ell(v_h) = \int_\Omega \nabla[i_h u - u] \cdot \nabla v_h \,\mathrm{d}x$$

for all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$. We have $\|\mathscr{C}_h(u)\|_{\mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)'} \le ch\|D^2 u\|$ if $u \in H^2(\Omega) \cap H^1_{\mathrm{D}}(\Omega)$ and $i_h u = \mathscr{I}_h u$ is the nodal interpolant of u. If $z_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ and $b_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)'$ are such that

$$\int_\Omega \nabla z_h \cdot \nabla v_h \,\mathrm{d}x = b_h(v_h)$$

for all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$, then the choice of $v_h = z_h$ shows the discrete stability estimate $\|\nabla z_h\| \le \|b_h\|_{\mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)'}$. Therefore, Theorem 4.3 implies the error estimate

$$\|\nabla(u_h - \mathscr{I}_h u)\|_{L^2(\Omega)} \le ch\|D^2 u\|_{L^2(\Omega)}.$$

(ii) Let $u_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ and define

$$\mathscr{R}_h(u_h; v) = F(u_h)[v] - \ell(v) = \int_\Omega \nabla u_h \cdot \nabla v \,\mathrm{d}x - \int_\Omega f v \,\mathrm{d}x - \int_{\Gamma_{\mathrm{N}}} g v \,\mathrm{d}s$$

for all $v \in H^1_{\mathrm{D}}(\Omega)$. Noting the stability estimate $\|\nabla z\| \le \|b\|_{X'}$ for $z \in H^1_{\mathrm{D}}(\Omega)$ and $b \in H^1_{\mathrm{D}}(\Omega)'$ with

$$\int_\Omega \nabla z \cdot \nabla v \,\mathrm{d}x = b(v)$$

for all $v \in H^1_{\mathrm{D}}(\Omega)$, Theorem 4.4 implies the error estimate

$$\|\nabla(u - u_h)\|_{L^2(\Omega)} \le \|\mathscr{R}_h(u_h)\|_{X'}.$$

If u_h satisfies $F(u_h)[v_h] = 0$ for all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$, we have the Galerkin orthogonality $F(u - u_h)[v_h] = 0$ for all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ and $\|\mathscr{R}_h(u_h)\|_{X'} \le c\eta(u_h)$ with a computable quantity $\eta(u_h)$, cf. Theorem 3.6.

The concepts can be generalized to the class of strongly monotone operators.

Definition 4.2 The operator $F : X \to X'$ is called *strongly monotone* if there exists an increasing bijection $\chi : [0, \infty) \to [0, \infty)$ with

$$\chi(\|u - v\|_X) \le \frac{\langle F(u) - F(v), u - v\rangle_X}{\|u - v\|_X}$$

for all $u, v \in X$.

We consider a conforming discretization of a strongly monotone problem in the following theorem.

Theorem 4.5 (Monotone problems) *Assume that $u \in X$ and $u_h \in X_h$ satisfy*

$$F(u)[v] = \ell(v), \qquad F(u_h)[v_h] = \ell(v_h)$$

for all $v \in X$ and $v_h \in X_h$, respectively, and let $\mathscr{C}_h(u)$ and $\mathscr{R}_h(u_h)$ for an interpolation operator i_h be defined by

$$\mathscr{C}_h(u; v_h) = F(i_h u)[v_h] - \ell(v_h), \qquad \mathscr{R}_h(u_h; v) = F(u_h)[v] - \ell(v)$$

for all $v_h \in X_h$ and $v \in X$, respectively. Then we have the a priori and a posteriori error estimates

$$\chi(\|i_h u - u_h\|_X) \le \|\mathscr{C}_h(u)\|_{X_h'}, \quad \chi(\|u - u_h\|_X) \le \|\mathscr{R}_h(u_h)\|_{X'}.$$

Proof We have

$$\|i_h u - u_h\|_X \, \chi(\|i_h u - u_h\|_X) \le \langle F(i_h u) - F(u_h), i_h u - u_h\rangle = \mathscr{C}_h(u; i_h u - u_h)$$

and

$$\|u - u_h\|_X \, \chi(\|u - u_h\|_X) \le \langle F(u) - F(u_h), u - u_h\rangle = -\mathscr{R}_h(u_h; u - u_h).$$

Dividing by $\|i_h u - u_h\|_X$ and $\|u - u_h\|_X$, respectively, yields the estimates. □

Example 4.13 (*p-Laplacian*) The p-Laplacian $-\operatorname{div}(|\nabla u|^{p-2}\nabla u)$ is identified with the functional $F : W^{1,p}_{\mathrm{D}}(\Omega) \to W^{1,p}_{\mathrm{D}}(\Omega)'$ defined by

$$F(u)[v] = \int_\Omega |\nabla u|^{p-2}\nabla u \cdot \nabla v \,\mathrm{d}x$$

for $u, v \in W^{1,p}_{\mathrm{D}}(\Omega)$. The functional F is the Fréchet derivative $F = DI$ of

$$I(u) = \frac{1}{p}\int_\Omega |\nabla u|^p \,\mathrm{d}x.$$

If $p \ge 2$, then F is monotone with $\chi(s) = \alpha s^{p-1}$ for all $s \ge 0$ and some $\alpha > 0$. The functional is locally Lipschitz continuous in the sense that

$$\|F(u) - F(v)\|_{W^{1,p}_{\mathrm{D}}(\Omega)'} \le M(\|\nabla u\|_{L^p(\Omega)} + \|\nabla v\|_{L^p(\Omega)})^{p-2}\|\nabla(u - v)\|_{L^p(\Omega)}$$

for a constant $M \in \mathbb{R}$ and $u, v \in W^{1,p}_{\mathrm{D}}(\Omega)$. This estimate implies the consistency of conforming discretizations, e.g., with $\mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$, and we obtain the error estimate

$$\alpha\|\nabla(i_h u - u_h)\|^{p-1}_{L^p(\Omega)} \le M\|\nabla(u - i_h u)\|_{L^p(\Omega)};$$

thus $\|\nabla(u - u_h)\|_{L^p(\Omega)} \le ch^{1/(p-1)}$ if $u \in W^{2,p}(\Omega) \cap W^{1,p}_{\mathrm{D}}(\Omega)$.

If the operator F fails to be monotone but has a regular Fréchet derivative in the neighborhood of a solution, then a local error estimate follows from the implicit function theorem. For ease of presentation and without loss of generality, we consider the homogeneous problem $F(u) = 0$.

Theorem 4.6 (Local error estimate [10]) *Suppose that $F : X \to X'$ is continuous and $u \in X$ satisfies $F(u) = 0$. Assume that there exist constants $c_1, c_2, c_3, \varepsilon > 0$ with $c_2 < c_1$ such that*

$$\|F(u) - F(v)\|_{X'} \le c_0 \|u - v\|_X,$$
$$\|DF(v)^{-1}\|_{L(X',X)} \le c_1^{-1},$$
$$\|DF(v) - DF(w)\|_{L(X,X')} \le c_2 \|v - w\|_X$$

for all $v, w \in B_\varepsilon(u)$. Let $i_h u \in X_h$ be an interpolant of u such that $c_0\|i_h u - u\|_X \le (c_1 - c_2)\varepsilon$. Then there exists a unique $u_h \in X_h$ with $F(u_h) = 0$ and $\|u - u_h\|_X \le \varepsilon$.

Proof The assumptions of the theorem imply that

$$\|F(i_h u)\|_{X'} = \|F(i_h u) - F(u)\|_{X'} \le c_0 \|u - i_h u\|_X.$$

A quantitative version of the implicit function theorem, cf. [2], implies the existence of a unique $u_h \in X_h$ with the asserted properties. □

Example 4.14 (*Semilinear diffusion*) The theorem implies error estimates for the approximation of the semilinear equation

$$-\Delta u + f(u) = 0 \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega,$$

provided that f' and a solution $u \in H_0^1(\Omega)$ are such that the operator $-\Delta + f'(v)\mathrm{id}$ is invertible for all $v \in B_\varepsilon(u)$ for some $\varepsilon > 0$. It is sufficient for this that $f' > -c_P^{-2}$ with the smallest constant $c_P > 0$, such that $\|w\| \le c_P \|\nabla w\|$ for all $w \in H_0^1(\Omega)$.

The following proposition generalizes the Lax–Milgram and the Céa lemma to bilinear forms that are not elliptic.

Proposition 4.2 (Generalized Lax–Milgram and Céa lemma [1, 13]) *Let X, Y be Hilbert spaces, $a : X \times Y \to \mathbb{R}$ a continuous bilinear form with continuity constant M, and $\ell \in Y'$. Assume that there exists $\alpha > 0$ such that*

$$\sup_{v \in Y \setminus \{0\}} \frac{a(u, v)}{\|v\|_Y} \ge \alpha \|u\|_X$$

for all $u \in X$ and that for all $v \in Y \setminus \{0\}$, there exists $u \in Y$ with $a(u, v) \neq 0$. Then there exists a unique $u \in X$ with

$$a(u, v) = \ell(v)$$

for all $v \in Y$ and $\|u\|_X \le \alpha^{-1}\|\ell\|_{Y'}$. If $X_h \subset X$ and $Y_h \subset Y$ are such that the above conditions are satisfied with X and Y replaced by X_h and Y_h, respectively, then there *exists a unique* $u_h \in X_h$ *with*

$$a(u_h, v_h) = \ell(v_h)$$

for all $v_h \in Y_h$*, and we have*

$$\|u - u_h\|_X \le (1 + \alpha^{-1} M) \inf_{w_h \in X_h} \|u - w_h\|_X.$$

Proof Identifying the bilinear form a with the operator $A : X \to Y'$, we see that A is injective, i.e., $Au = 0$ for $u \in X$ implies $u = 0$. Noting that

$$\alpha \|u_j - u_k\|_X \le \sup_{v \in Y \setminus \{0\}} \frac{\langle A(u_j - u_k), v\rangle}{\|v\|_Y} \le \|Au_j - Au_k\|_{Y'}$$

proves that the range of A is closed. If $v \in Y$ is such that $\langle Au, v\rangle = 0$ for all $u \in X$, then the assumptions imply $v = 0$. Hence, the closed range theorem yields that the range of A is Y' and it follows that A is bijective, i.e., there exists a unique $u \in X$ with $Au = \ell$. The estimate for $\|u\|_X$ is an immediate consequence of the assumptions. The same arguments show that the operator $A_h : X_h \to Y_h'$ is an isomorphism and hence there exists a unique $u_h \in X_h$ with the asserted properties. Let $w_h \in X_h$, and for every $v_h \in X_h$ define

$$\widetilde{\ell}(v_h) = a(u - w_h, v_h).$$

Then there exists a unique $z_h \in X_h$ with $a(z_h, v_h) = \widetilde{\ell}(v_h)$ and $\|z_h\|_X \le \alpha^{-1} \|\widetilde{\ell}\|_{Y_h'}$. Since $a(u_h, v_h) = a(u, v_h)$ it follows that $z_h = u_h - w_h$, and hence

$$\|u_h - w_h\|_X \le \alpha^{-1} M \|u - w_h\|.$$

The triangle inequality implies the asserted estimate. □

Example 4.15 (*Helmholtz equation*) Let $\omega \in \mathbb{R}$ and $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ be for $u, v \in H_0^1(\Omega)$ defined by

$$a(u, v) = (\nabla u, \nabla v) - \omega^2 (u, v),$$

which corresponds to the partial differential equation $-\Delta u - \omega^2 u = f$ in Ω with boundary condition $u|_{\partial\Omega} = 0$. If ω^2 is not an eigenvalue of $-\Delta$, then a satisfies the conditions of the proposition. To prove this, note that $(-\Delta)^{-1} : L^2(\Omega) \to H_0^1(\Omega) \subset L^2(\Omega)$ is selfadjoint and compact with trivial kernel, so that there exists a complete orthonormal system $(u_j)_{j\in\mathbb{N}} \subset L^2(\Omega)$ of eigenfunctions of $(-\Delta)^{-1}$, i.e., for every $j \in \mathbb{N}$ we have $-\Delta u_j = \lambda_j u_j$ with positive eigenvalues $(\lambda_j)_{j\in\mathbb{N}}$ that do not accumulate at zero. We have $\lambda_j^{-1}(\nabla u_j, \nabla u_k) = (u_j, u_k) = \delta_{jk}$ for all $j, k \in \mathbb{N}$. Given $u = \sum_{j\in\mathbb{N}} \alpha_j u_j \in H_0^1(\Omega)$, define $v = \sum_{j\in\mathbb{N}} \sigma_j \alpha_j u_j$ with $\sigma_j = \operatorname{sign}(\|\nabla u_j\|^2 - \omega^2 \|u_j\|^2)$. Then

$$a(u, v) = \sum_{j\in\mathbb{N}} \sigma_j \alpha_j^2 \big(\|\nabla u_j\|^2 - \omega^2 \|u_j\|^2\big) \ge \min_{j\in\mathbb{N}} \frac{|\lambda_j - \omega^2|}{\lambda_j} \|\nabla u\|^2$$

and with $\|\nabla u\| = \|\nabla v\|$, we deduce that

$$\sup_{v \in H_0^1(\Omega)} \frac{a(u, v)}{\|\nabla v\|} \geq c_H \|\nabla u\|.$$

The second condition of the proposition is a direct consequence of the requirement that ω^2 is not an eigenvalue of $-\Delta$.

Remark 4.5 Proposition 4.2 is important for the analysis of saddle-point problems; the seminal paper [7] provides conditions that imply the assumptions of the proposition.

4.2.3 Abstract Subdifferential Flow

The subdifferential flow of a convex and lower semicontinuous functional $I : H \to \mathbb{R} \cup \{+\infty\}$ arises as an evolutionary model in applications, and can be used as a basis for numerical schemes to minimize I. The corresponding differential equation seeks $u : [0, T] \to H$, such that $u(0) = u_0$ and

$$\partial_t u \in -\partial I(u),$$

i.e., $u(0) = u_0$ and

$$(-\partial_t u, v - u)_H + I(u) \leq I(v)$$

for almost every $t \in [0, T]$ and every $v \in H$. An implicit discretization of this nonlinear evolution equation is equivalent to a sequence of minimization problems involving a quadratic term. We recall that $d_t u^k = (u^k - u^{k-1})/\tau$ denotes the backward difference quotient.

Theorem 4.7 (Semidiscrete scheme [15, 17]) *Assume that $I \geq 0$ and for $u^0 \in H$ let $(u^k)_{k=1,\dots,K} \subset H$ be minimizers for*

$$I_\tau^k(w) = \frac{1}{2\tau} \|w - u^{k-1}\|_H^2 + I(w)$$

for $k = 1, 2, \dots, K$. For $L = 1, 2, \dots, K$, we have

$$I(u^L) + \tau \sum_{k=1}^{L} \|d_t u^k\|_H^2 \leq I(u^0).$$

With the computable quantities

$$\mathscr{E}_k = -\tau \|d_t u^k\|_H^2 - I(u^k) + I(u^{k-1})$$

and the affine interpolant $\widehat{u}_\tau : [0,T] \to H$ of the sequence $(u^k)_{k=0,\dots,K}$ we have the a posteriori error estimate

$$\max_{t\in[0,T]} \|u - \widehat{u}\|_H^2 \le \|u_0 - u^0\|_H^2 + \tau \sum_{k=1}^{L} \mathscr{E}_k.$$

We have the a priori error estimate

$$\max_{k=0,\dots,K} \|u(t_k) - u^k\|_H^2 \le \|u_0 - u^0\|_H^2 + \tau I(u^0),$$

and under the condition $\partial I(u^0) \neq \emptyset$, the improved variant

$$\max_{k=0,\dots,K} \|u(t_k) - u^k\|_H^2 \le \|u_0 - u^0\|_H^2 + \tau^2 \|\partial^o I(u^0)\|_H^2,$$

where $\partial^o I(u^0) \in H$ denotes the element of minimal norm in $\partial I(u^0)$.

Proof The direct method in the calculus of variations yields that for $k = 1, 2, \dots, K$, there exists a unique minimizer $u^k \in H$ for I_τ^k, and we have $d_t u^k \in -\partial I(u^k)$, i.e.,

$$(-d_t u^k, v - u^k)_H + I(u^k) \le I(v)$$

for all $v \in H$; the choice of $v = u^{k-1}$ implies that

$$-\mathscr{E}_k = \tau \|d_t u^k\|_H^2 + I(u^k) - I(u^{k-1}) \le 0$$

with $0 \le \mathscr{E}_k \le -\tau d_t I(u^k)$. A summation over $k = 1, 2, \dots, L$ yields the asserted stability estimate. If $\widehat{u}_\tau$ is the piecewise affine interpolant of $(u^k)_{k=0,\dots,K}$ associated to the time steps $t_k = k\tau$, $k = 0, 1, \dots, K$, and u_τ^+ is such that $u_\tau^+|_{(t_{k-1},t_k)} = u^k$ for $k = 1, 2, \dots$ and $t_k = k\tau$, then we have

$$(-\partial_t \widehat{u}_\tau, v - u_\tau^+)_H + I(u_\tau^+) \le I(v)$$

for almost every $t \in [0,T]$ and all $v \in H$. In introducing

$$\mathscr{C}_\tau(t) = (-\partial_t \widehat{u}_\tau, u_\tau^+ - \widehat{u}_\tau)_H - I(u_\tau^+) + I(\widehat{u}_\tau)$$

we have

$$(-\partial_t \widehat{u}_\tau, v - \widehat{u}_\tau)_H + I(\widehat{u}_\tau) \le I(v) + \mathscr{C}_\tau(t).$$

The choice of $v = u$ in this inequality and $v = \widehat{u}_\tau$ in the continuous evolution equation yield

$$\frac{d}{dt}\frac{1}{2}\|u - \widehat{u}\|_H^2 = (-\partial_t[u - \widehat{u}_\tau], \widehat{u}_\tau - u)_H \le \mathscr{C}_\tau(t).$$

Noting $\widehat{u}_\tau - u_\tau^+ = (t - t_k)\partial_t \widehat{u}_\tau$ for $t \in (t_{k-1}, t_k)$ and using the convexity of I, i.e.,

$$I(\widehat{u}_\tau) \leq \frac{t_k - t}{\tau} I(u^{k-1}) + \frac{t - t_{k-1}}{\tau} I(u^k),$$

we verify for $t \in (t_{k-1}, t_k)$ using $u_\tau^+ = u^k$ that

$$\mathscr{C}_\tau(t) \leq (t - t_k)\|\partial_t \widehat{u}_\tau\|_H^2 - I(u_\tau^+) + \frac{t_k - t}{\tau} I(u^{k-1}) + \frac{t - t_{k-1}}{\tau} I(u^k) = \frac{t_k - t}{\tau}\mathscr{E}_k.$$

With $\mathscr{E}_k \leq -\tau d_t I(u^k)$ and $I \geq 0$ we deduce that

$$\int_0^{t_L} \mathscr{C}_\tau(t)\,\mathrm{d}t \leq \tau \sum_{k=1}^{L} \mathscr{E}_k \leq -\tau^2 \sum_{k=1}^{L} d_t I(u^k) = -\tau\big(I(u^L) - I(u^0)\big) \leq \tau I(u^0),$$

which implies the a posteriori and the first a priori error estimate. Assume that $\partial I(u^0) \neq \emptyset$ and define $u^{-1} \in H$ so that $d_t u^0 = (u^0 - u^{-1})/\tau = -\partial^o I(u^0)$, i.e., the discrete evolution equation also holds for $k = 0$,

$$(-d_t u^0, v - u^0)_H + I(u^0) \leq I(v)$$

for all $v \in H$. Choosing $v = u^k$ in the equation for $d_t u^{k-1}$, $k = 1, 2, \dots, K$, we observe that

$$(-d_t u^{k-1}, u^k - u^{k-1})_H + I(u^{k-1}) \leq I(u^k),$$

i.e., $-\tau d_t I(u^k) \leq \tau(d_t u^k, d_t u^{k-1})_H$, and it follows that

$$\begin{aligned}\mathscr{E}_k &= -\tau(d_t u^k, d_t u^k)_H - \tau d_t I(u^k) \leq -\tau(d_t u^k, d_t u^k)_H + \tau(d_t u^{k-1}, d_t u^k)_H \\ &= -\tau^2(d_t^2 u^k, d_t u^k)_H = -\tau^2 \frac{d_t}{2}\|d_t u^k\|_H^2 - \frac{\tau^3}{2}\|d_t^2 u^k\|_H^2 \leq -\tau^2 \frac{d_t}{2}\|d_t u^k\|_H^2.\end{aligned}$$

This implies that

$$\int_0^{t_L} \mathscr{C}_\tau(t)\,\mathrm{d}t \leq \tau \sum_{k=1}^{L} \mathscr{E}_k \leq \frac{\tau^2}{2}\|d_t u^0\|_H^2 = \frac{\tau^2}{2}\|\partial^o I(u^0)\|_H^2,$$

which proves the improved a priori error estimate. □

Remarks 4.6 (i) The condition $\partial I(u^0) \neq \emptyset$ is restrictive in many applications.
(ii) Subdifferential flows $\partial_t u \in -\partial I(u)$, i.e., $Lu \ni 0$ for $Lu = \partial_t u + v$ with $v \in \partial I(u)$, and with a convex functional $I : H \to \mathbb{R} \cup \{+\infty\}$ define monotone

problems in the sense that

$$\begin{aligned}\big(Lu_1 - Lu_2, u_1 - u_2\big)_H &= (\partial_t(u_1 - u_2) + (v_1 - v_2), u_1 - u_2\big)_H \\ &\geq \big(\partial_t(u_1 - u_2), u_1 - u_2\big)_H = \frac{1}{2}\frac{d}{dt}\|u_1 - u_2\|_H^2\end{aligned}$$

for u_1, u_2 and v_1, v_2 with $v_i \in \partial I(u_i)$, $i = 1, 2$.
(iii) If $I : H \to \mathbb{R} \cup \{+\infty\}$ is strongly monotone in the sense that $(u_1 - u_2, v_1 - v_2)_H \geq \alpha\|u_1 - u_2\|_H^2$ whenever $v_\ell \in \partial I(u_\ell)$, $\ell = 1, 2$, and if there exists a solution $\overline{u} \in H$ of the stationary inclusion $\overline{v} = 0 \in \partial I(\overline{u})$, then we have $u(t) \to \overline{u}$ as $t \to \infty$. A proof follows from the estimate

$$\frac{1}{2}\frac{d}{dt}\|u - \overline{u}\|_H^2 = -(v - \overline{v}, u - \overline{u})_H \leq -\alpha\|u - \overline{u}\|_H^2,$$

where $v = -\partial_t u \in \partial I(u)$, and an application of Gronwall's lemma.

4.2.4 Weak Continuity Methods

Let $(u_h)_{h>0} \subset X$ be a bounded sequence in the reflexive, separable Banach space X such that there exists a weak limit $u \in X$ of a subsequence that is not relabeled, i.e., we have $u_h \rightharpoonup u$ as $h \to 0$. For an operator $F : X \to X'$, we define the sequence $(\xi_h)_{h>0} \subset X'$ through $\xi_h = F(u_h)$, and if the sequence is bounded in X', then there exists $\xi \in X'$, such that for a further subsequence $(\xi_h)_{h>0}$ which again is not relabeled, we have $\xi_h \rightharpoonup^* \xi$. The important question is now whether we have weak continuity in the sense that

$$F(u) = \xi.$$

Notice that weak continuity is a strictly stronger notion of continuity than strong continuity. For partial differential equations, this property is called *weak precompactness* of the solution set of the homogeneous equation, i.e., if $(u_j)_{j\in\mathbb{N}}$ is a sequence with $F(u_j) = 0$ for all $j \in \mathbb{N}$ and $u_j \rightharpoonup u$ as $j \to \infty$ then we may deduce that $F(u) = 0$. Such implications may also be regarded as properties of weak stability since they imply that if $F(u_j) = r_j$ with $\|r_j\|_{X'} \leq \varepsilon_j$ and $\varepsilon_j \to 0$ as $j \to \infty$, then we have $F(u) = 0$ for every accumulation point of the sequence $(u_j)_{j\in\mathbb{N}}$.

Theorem 4.8 (Discrete compactness) *For every $h > 0$ let $u_h \in X_h$ solve $F_h(u_h) = 0$. Assume that $F_h(u_h) \in X'$ with $\|F(u_h)\|_{X'} \leq c$ for all $h > 0$ and F is weakly continuous on X, i.e., $F(u_j)[v] \to F(u)[v]$ for all $v \in X$ whenever $u_j \rightharpoonup u$ in X. Suppose that for every bounded sequence $(w_h)_{h>0} \subset X$ with $w_h \in X_h$ for all $h > 0$, we have*

$$\|F(w_h) - F_h(w_h)\|_{X_h'} \to 0$$

as $h \to 0$ and $(X_h)_{h>0}$ is dense in X with respect to strong convergence. If $(u_h)_{h>0} \subset X$ is bounded, then there exists a subsequence $(u_{h'})_{h'>0}$ and $u \in X$ such that $u_h \rightharpoonup u$ in X and $F(u) = 0$.

Proof After extraction of a subsequence, we may assume that $u_h \rightharpoonup u$ in X as $h \to 0$ for some $u \in X$. Fixing $v \in X$ and using $F_h(u_h)[v_h] = 0$ for every $v_h \in X_h$, we have

$$F(u_h)[v] = F(u_h)[v - v_h] + F(u_h)[v_h] - F_h(u_h)[v_h].$$

For a sequence $(v_h)_{h>0} \subset X$ with $v_h \in X_h$ for every $h > 0$ and $v_h \to v$ in X, we find that

$$|F(u_h)[v - v_h]| \le \|F(u_h)\|_{X'} \|v - v_h\|_X \to 0$$

as $h \to 0$. The sequences $(u_h)_{h>0}$ and $(v_h)_{h>0}$ are bounded in X and thus

$$|F(u_h)[v_h] - F_h(u_h)[v_h]| \le \|F(u_h) - F_h(u_h)\|_{X_h'} \|v_h\|_X \to 0$$

as $h \to 0$. Together with the weak continuity of F we find that

$$F(u)[v] = \lim_{h \to 0} F(u_h)[v] = 0.$$

Since $v \in X$ was arbitrary this proves the theorem. □

The crucial part in the theorem is the weak continuity of the operator F. We include an example of an operator related to a constrained nonlinear partial differential equation that fulfills this requirement.

Example 4.16 (*Harmonic maps*) Let $(u_j)_{j \in \mathbb{N}} \subset H^1(\Omega; \mathbb{R}^3)$ be a bounded sequence such that $|u_j(x)| = 1$ for all $j \in \mathbb{N}$ and almost every $x \in \Omega$. Assume that for every $j \in \mathbb{N}$ and all $v \in H^1(\Omega; \mathbb{R}^3) \cap L^\infty(\Omega; \mathbb{R}^3)$, we have

$$F(u_j)[v] = \int_\Omega \nabla u_j \cdot \nabla v \,\mathrm{d}x - \int_\Omega |\nabla u_j|^2 u_j \cdot v \,\mathrm{d}x = 0.$$

The choice of $v = u_j \times w$ shows that we have

$$\widetilde{F}(u_j)[w] = \int_\Omega \nabla u_j \cdot \nabla(u_j \times w) \,\mathrm{d}x = 0$$

for all $w \in H^1(\Omega; \mathbb{R}^3) \cap L^\infty(\Omega; \mathbb{R}^3)$. Using $\partial_k u_j \cdot \partial_k(u_j \times w) = \partial_k u_j \cdot (u_j \times \partial_k w)$ for $k = 1, 2, \dots, d$, we find that

$$\widetilde{F}(u_j)[w] = \sum_{k=1}^d \int_\Omega \partial_k u_j \cdot (u_j \times \partial_k w) \,\mathrm{d}x = 0.$$

If $u_j \rightharpoonup u$ in $H^1_{\mathrm{D}}(\Omega;\mathbb{R}^3)$, then $u_j \to u$ in $L^2(\Omega;\mathbb{R}^3)$ and thus, for every fixed $w \in C^\infty(\overline{\Omega};\mathbb{R}^3)$, we can pass to the limit and find that

$$\widetilde{F}(u)[w] = 0.$$

Since up to a subsequence we have $u_j(x) \to u(x)$ for almost every $x \in \Omega$, we verify that $|u(x)| = 1$ for almost every $x \in \Omega$. A density result shows that this holds for all $w \in H^1(\Omega;\mathbb{R}^3) \cap L^\infty(\Omega;\mathbb{R}^3)$. Reversing the above argument by choosing $w = u \times v$ and employing the identity $a \times (b \times c) = (b \cdot a)c - (c \cdot a)b$ shows that $F(u)[v] = 0$ for all $v \in H^1(\Omega;\mathbb{R}^3) \cap L^\infty(\Omega;\mathbb{R}^3)$.

A general concept for weak continuity is based on the notion of pseudomonotonicity.

Example 4.17 (*Pseudomonotone operators*) The operator $F : X \to X'$ is a *pseudomonotone* operator if it is bounded, i.e., $\|F(u)\|_{X'} \le c(1 + \|u\|_X^s)$ for some $s \ge 0$, and whenever $u_j \rightharpoonup u$ in X, we have the implication that

$$\limsup_{j\to\infty} F(u_j)[u_j - u] \le 0 \quad \Longrightarrow \quad F(u)[u - v] \le \liminf_{j\to\infty} F(u_j)[u_j - v].$$

For such an operator we have that if $F(u_h)[v_h] = \ell(v_h)$ for all $v_h \in X_h$ with a strongly dense family of subspaces $(X_h)_{h>0}$ and $u_h \rightharpoonup u$ as $h \to 0$, then $F(u) = \ell$. To verify this, let $v \in X$ and $(v_h)_{h>0}$ with $v_h \in X_h$ such that $v_h \to u$ and note that

$$\begin{aligned}\limsup_{h\to 0} F(u_h)[u_h - u] &= \limsup_{h\to 0} F(u_h)[u_h - v_h] + F(u_h)[v_h - u] \\ &= \limsup_{h\to 0} \ell(u_h - v_h) + F(u_h)[v_h - u] = 0.\end{aligned}$$

Pseudomonotonicity yields for every $v_{h'} \in \cup_{h>0} X_h$ that

$$F(u)[u - v_{h'}] \le \liminf_{h\to 0} F(u_h)[u_h - v_{h'}] = \lim_{h\to 0} \ell(u_h - v_{h'}) = \ell(u - v_{h'}).$$

With the density of $(X_h)_{h>0}$ in X, we conclude that $F(u)[u - v] \le \ell(u - v)$ for all $v \in X$ and with $v = u \pm w$, we find that $F(u)[w] = \ell(w)$ for all $w \in X$.

Remarks 4.7 (i) Radially continuous bounded operators are pseudomonotone. Here, radial continuity means that $t \mapsto F(u + tv)[v]$ is continuous for $t \in \mathbb{R}$ and all $u, v \in X$. These operators allow us to apply Minty's trick to deduce from the inequality $\ell(u - v) - F(v)[u - v] \ge 0$ for all $v \in X$ that $F(u) = \ell$. To prove this implication, note that with $v = u + \varepsilon w$, we find that $\ell(w) - F(u + \varepsilon w)[w] \le 0$ and by radial continuity for $\varepsilon \to 0$, it follows that $\ell(w) - F(u)[w] \le 0$ and hence $F(u) = \ell$.

(ii) Pseudomonotone operators are often of the form $F = F_1 + F_2$ with a monotone operator F_1 and a weakly continuous operator F_2, e.g., a lower-order term described by F_2.

Example 4.18 (*Quasilinear diffusion*) The concept of pseudomonotonicity applies to the quasilinear elliptic equation

$$-\operatorname{div}\left(|\nabla u|^{p-2}\nabla u\right) + g(u) = f \text{ in } \Omega, \quad u|_{\partial\Omega} = 0 \text{ on } \partial\Omega,$$

with $g \in C(\mathbb{R})$ such that $|g(s)| \leq c(1 + |s|^{r-1})$ and $1 < p < d, r < dp/(d-p)$.

4.3 Solution of Discrete Problems

We discuss in this section the practical solution of discretized minimization problems of the form

$$\text{Minimize } I_h(u_h) = \int_\Omega W(\nabla u_h) + g(u_h)\,\mathrm{d}x \text{ among } u_h \in \mathscr{A}_h.$$

In particular, we investigate four model situations with smooth and nonsmooth integrands and smooth and nonsmooth constraints included in $\mathscr{A}$. The iterative algorithms are based on an approximate solution of the discrete Euler–Lagrange equations. More general results can be found in the textbooks [4, 12].

4.3.1 Smooth, Unconstrained Minimization

Suppose that

$$\mathscr{A}_h = \{u_h \in \mathscr{S}^1(\mathscr{T}_h)^m : u_h|_{\Gamma_\mathrm{D}} = u_{\mathrm{D},h}\}$$

and I_h is defined as above with functions $W \in C^1(\mathbb{R}^{m\times d})$ and $g \in C^1(\mathbb{R}^m)$. The case $\Gamma_\mathrm{D} = \emptyset$ is not generally excluded in the following. A necessary condition for a minimizer $u_h \in \mathscr{A}_h$ is that for all $v_h \in \mathscr{S}^1_\mathrm{D}(\mathscr{T}_h)^m$, we have

$$F_h(u_h)[v_h] = \int_\Omega DW(\nabla u_h) \cdot \nabla v_h + Dg(u_h) \cdot v_h\,\mathrm{d}x = 0.$$

Steepest descent methods successively lower the energy by minimizing in descent directions defined through an appropriate gradient.

Algorithm 4.1 (*Descent method*) *Let* $(\cdot,\cdot)_H$ *be a scalar product on* $\mathscr{S}^1_\mathrm{D}(\mathscr{T}_h)^m$ *and* $\mu \in (0, 1/2)$. *Given* $u_h^0 \in \mathscr{A}_h$, *compute the sequence* $(u_h^j)_{j=0,1,\ldots}$ *via* $u_h^{j+1} = u_h^j + \alpha_j d_h^j$ *with* $d_h^j \in \mathscr{S}^1_\mathrm{D}(\mathscr{T}_h)^m$ *such that*

$$(d_h^j, v_h)_H = -F_h(u_h^j)[v_h]$$

for all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m$ and either the fixed step-size

$$\alpha_j = \tau$$

or the line-search minimum which seeks the maximal $\alpha_j \in \{2^{-\ell},\ \ell \in \mathbb{N}_0\}$ such that

$$I_h(u_h^j + \alpha_j d_h^j) \le I_h(u_h^j) - \mu\alpha_j \|d_h^j\|_H^2.$$

Stop the iteration if $\|\alpha_j d_h^j\|_H \le \varepsilon_{\mathrm{stop}}$.

Remarks 4.8 (i) Since I_h is continuously differentiable, the descent method decreases the energy in every step. This follows from

$$\frac{d}{d\alpha} I_h(u_h^j + \alpha d_h^j)\Big|_{\alpha=0} = DI_h(u_h^j)[d_h^j] = F_h(u_h^j)[d_h^j] = -\|d_h^j\|_H^2,$$

i.e., the continuous function $\varphi(\alpha) = I_h(u_h^j + \alpha d_h^j)$ is strictly decreasing for $\alpha \in [0, \delta]$. The existence of $\alpha_j > 0$ that satisfies the Armijo–Goldstein condition of Algorithm 4.1 follows from expanding

$$I_h(u_h^j + \alpha d_h^j) = I_h(u_h^j) - \alpha\|d_h^j\|_H^2 + \mathscr{O}(\alpha^2)$$

provided that W and g are sufficiently smooth so that $I_h \in C^2(X_h)$.
(ii) The scalar product $(\cdot,\cdot)_H$ acts like a preconditioner for F_h, i.e., we have $u_h^{j+1} = u_h^j - \tau X_H^{-1} F_h(u_h^j)$ with respect to an appropriate basis. In particular, the descent method may be regarded as a fixed-point iteration.
(iii) Larger step sizes are typically possible for implicit or semi-implicit versions of the descent method, i.e., by considering a fixed step-size and the modified equation

$$(d_h^j, v_h)_H + \widetilde{F}_h(u_h^j + \tau d_h^j, u_h^j)[v_h] = 0$$

for all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m$ and with a function $\widetilde{F}_h$ such that $\widetilde{F}_h(u_h, u_h) = F_h(u_h)$. If $F_h(u_h) = G_h(u_h) + T_h(u_h)$ with a linear or monotone operator G_h, then a natural choice is $\widetilde{F}_h(u_h, \widetilde{u}_h) = G_h(u_h) + T_h(\widetilde{u}_h)$. Generally, large time steps are possible when monotone terms are treated implicitly and antimonotone terms explicitly.
(iv) If $X_h = V_h \times W_h$ and $I_h(u_h) = J_h(\phi_h, \psi_h)$ is separately convex, i.e., the mappings $v_h \mapsto J_h(v_h, \psi_h)$ and $w_h \mapsto J_h(\phi_h, w_h)$ are convex for all $(\phi_h, \psi_h) \in V_h \times W_h$, a decoupled, semi-implicit gradient flow discretization is unconditionally stable. Given the initial $(\phi_h^0, \psi_h^0) \in V_h \times W_h$, consider the iteration

$$(d_t\phi_h^{j+1}, v_h)_{V_h} + \delta_1 J_h(\phi_h^{j+1}, \psi_h^j)[v_h] = 0,$$

$$(d_t\psi_h^{j+1}, w_h)_{W_h} + \delta_2 J_h(\phi_h^{j+1}, \psi_h^{j+1})[w_h] = 0,$$

where $\delta_1 J_h$ and $\delta_2 J_h$ denote the Fréchet derivatives of J_h with respect to the first and second argument, respectively. The choices $v_h = d_t\phi_h^{j+1}$, $w_h = d_t\psi_h^{j+1}$ and the separate convexity of J lead to

$$\begin{aligned}\|d_t\phi_h^{j+1}\|_{V_h}^2 + \|d_t\psi_h^{j+1}\|_{W_h}^2 &= -\delta_1 J_h(\phi_h^{j+1}, \psi_h^j)[d_t\phi_h^{j+1}] \\ &\quad - \delta_2 J_h(\phi_h^{j+1}, \psi_h^{j+1})[d_t\psi_h^{j+1}] \\ &\le \tau^{-1}\big(J_h(\phi_h^j, \psi_h^j) - J_h(\phi_h^{j+1}, \psi_h^j)\big) \\ &\quad + \tau^{-1}\big(J_h(\phi_h^{j+1}, \psi_h^j) - J_h(\phi_h^{j+1}, \psi_h^{j+1})\big) \\ &= -d_t J_h(\phi_h^{j+1}, \psi_h^{j+1}),\end{aligned}$$

which implies the unconditional stability of the scheme.

Theorem 4.9 (Convex functionals) *Assume that I_h is convex and bounded from below and F_h is Lipschitz continuous, i.e., there exists $c_F \ge 0$ such that*

$$\|F_h(w_h) - F_h(v_h)\|_{X_h'} \le c_F \|w_h - v_h\|_X$$

for all $w_h, v_h \in X_h$. Let $c_h > 0$ be such that $\|v_h\|_X \le c_h\|v_h\|_H$ for all $v_h \in X_h$. Then the steepest descent method with fixed step-size $\tau > 0$ such that $\tau c_F c_h \le 1/2$ terminates within a finite number of iterations, and for all $J \ge 0$, we have

$$I_h(u_h^{J+1}) + (\tau/2)\sum_{j=0}^{J} \|d_h^j\|_H^2 \le I_h(u_h^0).$$

Proof The convexity of I_h implies that

$$F_h(u_h^{j+1})[u_h^{j+1} - u_h^j] + I_h(u_h^j) \ge I_h(u_h^{j+1}).$$

Using that $\tau d_h^j = u_h^{j+1} - u_h^j$ and choosing $v_h = \tau d_h^j$ in the discrete scheme leads to

$$\begin{aligned}I_h(u_h^{j+1}) - I_h(u_h^j) + \tau\|d_h^j\|_H^2 &\le (d_h^j, d_h^j)_H + \tau F_h(u_h^{j+1})[d_h^j] \\ &= (d_h^j, d_h^j)_H - F_h(u_h^j)[d_h^j] \\ &\quad + \tau\big(F_h(u_h^j) - F_h(u_h^{j+1})\big)[d_h^j] \\ &= \tau\big(F_h(u_h^j) - F_h(u_h^{j+1})\big)[d_h^j] \le c_F c_h \tau^2 \|d_h^j\|_H^2.\end{aligned}$$

Therefore, if $\tau c_F c_h \le 1/2$ we deduce the estimate from a summation over $j = 0, 1, \ldots, J$. The estimate implies that $d_h^j \to 0$ as $j \to \infty$ so that $\|\tau d_h^j\|_H \le \varepsilon_{\text{stop}}$ for j sufficiently large. □

Remarks 4.9 (i) The arguments of the proof of the theorem show that the implicit version of the descent method, defined by $(d_h^j, v_h)_H + F_h(u_h^j + \tau d_h^j)[v_h] = 0$ for every $v_h \in \mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h)$, is unconditionally convergent, but requires the solution of nonlinear systems of equations in every time step.
(ii) For nonconvex functionals, the iteration typically converges to a local minimum of I_h. Theoretically, the iteration may stop at a saddle point or local maximum.

To formulate the Newton method for solving the equation $F_h(u_h) = 0$ in X_h' we assume that $W \in C^2(\mathbb{R}^{m\times d})$ and $g \in C^2(\mathbb{R}^m)$. The Newton scheme may be regarded as an explicit descent method with a variable metric defined by the second variation of the energy functional I_h, i.e.,

$$DF_h(u_h)[w_h, v_h] = \int_\Omega D^2W(\nabla u_h)[\nabla w_h, \nabla v_h] + D^2 g(u_h)[w_h, v_h]\,\mathrm{d}x$$

for $u_h, v_h, w_h \in \mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h)^m$.

Algorithm 4.2 (*Newton method*) *Given $u_h^0 \in \mathscr{A}_h$, compute the sequence $(u_h^j)_{j=0,1,\dots}$ via $u_h^{j+1} = u_h^j + \alpha_j d_h^j$ with $d_h^j \in \mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h)^m$ such that*

$$DF_h(u_h^j)[d_h^j, v_h] = -F_h(u_h^j)[v_h]$$

for all $v_h \in \mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h)^m$ and $\alpha_j > 0$ with either the optimal step-size $\alpha_j = 1$, a fixed damping parameter $\alpha_j = \tau < 1$, or a line search minimum α_j as in Algorithm 4.1. Stop the iteration if $\|\alpha_j d_h^j\|_H \le \varepsilon_{\mathrm{stop}}$ for a norm $\|\cdot\|_H$ on $\mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h)^m$.

The convergence of the Newton iteration will be discussed in a more general context below in Sect. 4.3.3.

Remark 4.10 As opposed to the descent method, the Newton iteration can in general only be expected to converge locally. Under certain conditions the Newton scheme converges quadratically in a neighborhood of a solution. Optimal results can be obtained by combining the globally but slowly convergent descent method with the locally but rapidly convergent Newton method. Since the convergence of the Newton method is often difficult to establish and requires W and g to be sufficiently regular, developing globally convergent schemes is important to construct reliable numerical methods.

Example 4.19 For the approximation of minimal surfaces that are presented by graphs of functions over Ω, we consider

$$I_h(u_h) = \int_\Omega (1 + |\nabla u_h|^2)^{1/2}\,\mathrm{d}x$$

and note that for $u_h \in \mathscr{A}_h = \{v_h \in \mathscr{S}^1(\mathscr{T}_h) : v_h|_{\Gamma_\mathrm{D}} = u_{\mathrm{D},h}\}$ and $v_h, w_h \in \mathscr{S}^1_\mathrm{D}(\mathscr{T}_h)$ we have

$$F_h(u_h)[v_h] = \int_\Omega \frac{\nabla u_h \cdot \nabla v_h}{(1+|\nabla u_h|^2)^{1/2}}\,\mathrm{d}x$$

and

$$DF_h(u_h)[w_h, v_h] = \int_\Omega \frac{\nabla w_h \cdot \nabla v_h}{(1+|\nabla u_h|^2)^{1/2}} - \frac{\big(\nabla u_h \cdot \nabla v_h\big)\big(\nabla u_h \cdot \nabla w_h\big)}{(1+|\nabla u_h|^2)^{3/2}}\,\mathrm{d}x.$$

Figure 4.2 displays a combined MATLAB implementation of the Newton iteration and the descent method with line search. The Newton method fails to provide meaningful approximations for moderate perturbations of the nodal interpolant of the exact solution as a starting value.

4.3.2 Smooth Constrained Minimization

We next consider the case that the set of admissible functions includes a pointwise constraint, which is imposed at the nodes of a triangulation, i.e., for $G \in C(\mathbb{R}^m)$, we have

$$\mathscr{A}_h = \{u_h \in \mathscr{S}^1(\mathscr{T}_h)^m : u_h|_{\Gamma_\mathrm{D}} = u_{\mathrm{D},h},\ G\big(u_h(z)\big) = 0 \quad \text{for all } z \in \mathscr{N}_h\}.$$

The identity $G\big(u_h(z)\big) = 0$ for all $z \in \mathscr{N}_h$ is equivalent to the condition $\mathscr{I}_h G(u_h) = 0$. We always assume in the following that $\mathscr{A}_h \neq \emptyset$, i.e., that the function $u_{\mathrm{D},h}$ is compatible with the constraint. Moreover, we assume $G \in C^1(\mathbb{R}^m)$ with $DG(s) \neq 0$ for every $s \in M = G^{-1}(\{0\})$ so that $M \subset \mathbb{R}^m$ is an $(m-1)$-dimensional C^1-submanifold. The Euler–Lagrange equations of the discrete minimization problem

$$I_h(u_h) = \int_\Omega W(\nabla u_h)\,\mathrm{d}x$$

in the set of all functions $u_h \in \mathscr{A}_h$ can then be formulated as follows.

Proposition 4.3 (Optimality conditions) *The function $u_h \in \mathscr{A}_h$ is stationary for I_h in $\mathscr{A}_h$ if and only if*

$$F_h(u_h)[w_h] = 0$$

for all $w_h \in T_{u_h}\mathscr{A}_h$, where the discrete tangent space $T_{u_h}\mathscr{A}_h$ of $\mathscr{A}_h$ at u_h is defined by

$$T_{u_h}\mathscr{A}_h = \{w_h \in \mathscr{S}^1_\mathrm{D}(\mathscr{T}_h)^m : DG(u_h(z))w_h(z) = 0 \quad \text{forall}\, z \in \mathscr{N}_h \setminus \Gamma_\mathrm{D}\}.$$

```
function min_surf(red,scheme)
[c4n,n4e,Db,Nb] = triang_ring(red); nC = size(c4n,1);
[s,m,m_lumped] = fe_matrices(c4n,n4e); X_metric = s;
dNodes = unique(Db); fNodes = setdiff(1:nC,dNodes);
u = u_D(c4n); sd = zeros(nC,1);
pert = .01*(rand(nC,1)-.5); pert(dNodes) = 0; u = u+pert;
mu = 1/4; norm_corr = 1; eps_stop = 1e-4;
while norm_corr > eps_stop
    alpha = 1;
    if strcmp(scheme,'descent')
        [I_0,dI,¬] = energy(c4n,n4e,u);
        sd(fNodes) = -X_metric(fNodes,fNodes)\dI(fNodes);
        [I_alpha,¬] = energy(c4n,n4e,u+alpha*sd);
        armijo = I_alpha-I_0-mu*alpha*dI'*sd;
        while armijo > 0
            alpha = alpha/2;
            [I_alpha,¬] = energy(c4n,n4e,u+alpha*sd);
            armijo = I_alpha-I_0-mu*alpha*dI'*sd;
        end
    elseif strcmp(scheme,'newton')
        [¬,dI,d2I] = energy(c4n,n4e,u);
        sd(fNodes) = -d2I(fNodes,fNodes)\dI(fNodes);
    end
    u = u+alpha*sd; show_p1(c4n,n4e,Db,Nb,u);
    norm_corr = sqrt((alpha*sd)'*X_metric*(alpha*sd))
end

function [I,dI,d2I] = energy(c4n,n4e,u)
[nC,d] = size(c4n); nE = size(n4e,1);
ctr_max = (d+1)^2*nE; ctr = 0;
I1 = zeros(ctr_max,1); I2 = zeros(ctr_max,1);
X_d2I = zeros(ctr_max,1);
I = 0; dI = zeros(nC,1);
for j = 1:nE
    grads_T = [1,1,1;c4n(n4e(j,:),:)']\[0,0;eye(2)];
    vol_T = det([1,1,1;c4n(n4e(j,:),:)'])/2;
    du = (grads_T)'*u(n4e(j,:)); mod_du = norm(du);
    I = I+vol_T*(1+mod_du^2)^(1/2);
    P_loc = ((1+mod_du^2).*eye(d)-du*du')./((1+mod_du^2)^(3/2));
    for k = 1:d+1
        dI(n4e(j,k)) = dI(n4e(j,k))...
            +vol_T*grads_T(k,:)*du/(1+mod_du^2);
        for ell = 1:d+1
            ctr = ctr+1; I1(ctr) = n4e(j,k); I2(ctr) = n4e(j,ell);
            X_d2I(ctr) = vol_T*(P_loc*grads_T(k,:)')'...
                *grads_T(ell,:)';
        end
    end
end
d2I = sparse(I1,I2,X_d2I);

function val = u_D(x)
val = .5*(2-sqrt(sum(x.^2,2)));
```

Fig. 4.2 MATLAB routine for the computation of discrete minimal surfaces with the Newton and the steepest descent scheme

Proof We let $\varphi_h : (-\varepsilon, \varepsilon) \to \mathscr{A}_h$ be a continuously differentiable function with $\varphi_h(0) = u_h$. We then have that $w_h = \varphi_h'(0) \in T_{u_h}\mathscr{A}_h$ and

$$0 = \frac{d}{dt} I_h(\varphi_h(t))\Big|_{t=0} = DI_h(u_h)[w_h].$$

Conversely, for every $w_h \in T_{u_h}\mathscr{A}_h$ there exists a function $\varphi_h(t)$ as above. □

Remark 4.11 An equivalent characterization of stationary points is the existence of a Lagrange multiplier $\lambda_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ such that for all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m$, we have

$$F_h(u_h)[v_h] + (\lambda_h DG(u_h), v_h)_h = 0.$$

We propose the following descent scheme for the iterative solution of the constrained problem. It may be regarded as a semi-implicit discretization of an H-gradient flow. In particular, the problems that have to be solved at every step of the iteration are linear if F_h is linear.

Algorithm 4.3 (*Constrained descent method*) *Let* $(\cdot,\cdot)_H$ *be a scalar product on* $\mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m$ *and given* $u_h^0 \in \mathscr{A}_h$*, compute the sequence* $(u_h^j)_{j=0,1,\ldots}$ *via* $u_h^{j+1} = u_h^j + \tau d_h^j$ *with* $d_h^j \in T_{u_h^j}\mathscr{A}_h$ *such that*

$$(d_h^j, v_h)_H + F_h(u_h^j + \tau d_h^j)[v_h] = 0$$

for all $v_h \in T_{u_h^j}\mathscr{A}_h$*. Stop the iteration if* $\|d_h^j\|_H \le \varepsilon_{\mathrm{stop}}$.

Remark 4.12 If F_h is linear, then the solution of an iteration is equivalent to the solution of a linear system of equations of the form

$$\begin{bmatrix} X_H + \tau S & dG^\top \\ dG & 0 \end{bmatrix} \begin{bmatrix} D_h^j \\ \Lambda_h^j \end{bmatrix} = \begin{bmatrix} -SU_h^j \\ 0 \end{bmatrix},$$

where D_h^j, U_h^j, and Λ_h^j are vectors that contain the nodal values of the functions d_h^j, u_h^j, and λ_h^j, respectively, and X_H, S, and dG_h are matrices that represent the scalar product $(\cdot,\cdot)_H$, the bilinear form $F_h(u_h)[v_h]$, and the linearized constraint defined by DG.

The iterates $(u_h^j)_{j=0,1,\ldots}$ will in general not satisfy the constraint $\mathscr{I}_h G(u_h^j) = 0$ but under moderate conditions, the violation of the constraint is small. We recall the notation $\|v\|_h^2 = \int_\Omega \mathscr{I}_h[v^2]\,\mathrm{d}x$ for $v \in C(\overline{\Omega})$.

Theorem 4.10 (Constrained convex minimization) *Assume that* $G \in C^2(\mathbb{R}^m)$ *with* $\|D^2G\|_{L^\infty(\mathbb{R}^m)} \le c$*,* I_h *is convex,* $u_h^0 \in \mathscr{A}_h$*, and* $\|v_h\|_h \le c\|v_h\|_H$ *for all* $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m$*. For all* $J \ge 0$ *we have*

$$I_h(u_h^{J+1}) + \tau \sum_{j=0}^{J} \|d_h^j\|_H^2 \le I_h(u_h^0),$$

and for every $j = 1, 2, \ldots$, *the bound*

$$\|\mathscr{I}_h G(u_h^{j+1})\|_{L^1(\Omega)} \le c\tau I_h(u_h^0).$$

The algorithm terminates after a finite number of iterations.

Proof The convexity of I_h implies that

$$I_h(u_h^j + \tau d_h^j) + F_h(u_h + \tau d_h^j)[u_h^j - (u_h^j + \tau d_h^j)] \le I_h(u_h^j).$$

With the choice of $v_h = \tau d_h^j$ in the algorithm and the relation $u_h^{j+1} = u_h^j + \tau d_h^j$, this leads to

$$I_h(u_h^{j+1}) - I_h(u_h^j) \le \tau F_h(u_h^{j+1})[d_h^j] = -\tau \|d_h^j\|_H^2.$$

A summation over $j = 0, 1, \ldots, J$ proves the energy law. A Taylor expansion shows that for every $z \in \mathscr{N}_h \backslash \Gamma_{\mathrm{D}}$, we have for some $\xi_z^j \in \mathbb{R}^m$ that

$$G(u_h^{j+1}(z)) = G(u_h^j(z)) + \tau DG(u_h^j(z)) \cdot d_h^j(z) + \frac{\tau^2}{2} d_h^j(z)^\top D^2G(\xi_z^j) d_h^j(z).$$

Noting $DG(u_h^j(z)) \cdot d_h^j(z) = 0$ and $G(u_h^0(z)) = 0$, we deduce by induction that

$$G(u_h^{J+1}(z)) = \frac{\tau^2}{2} \sum_{j=0}^{J} d_h^j(z)^\top D^2G(\xi_z^j) d_h^j(z).$$

Since D^2G is uniformly bonded we have with $\beta_z = \int_\Omega \varphi_z \,\mathrm{d}x$ that

$$\begin{aligned}\|\mathscr{I}_h G(u_h^{j+1})\|_{L^1(\Omega)} \le \sum_{z \in \mathscr{N}} \beta_z |G(u_h^j(z))| &\le c\tau^2 \sum_{j=0}^{J} \sum_{z \in \mathscr{N}} \beta_z |d_h^j(z)|^2 \\ &= c\tau^2 \sum_{j=0}^{J} \|d_h^j\|_h^2.\end{aligned}$$

A combination with the energy law implies the bound for $\|\mathscr{I}_h G(u_h^{j+1})\|_{L^1(\Omega)}$. The convergence of the iteration follows from the convergence of the sum of norms of the correction vectors d_h^j. □

Remark 4.13 In order to satisfy the constraint exactly, the algorithm can be augmented by defining the new iterates through the projection

$$u_h^{j+1}(z) = \pi_M\big(u_h^j(z) + \tau d_h^j(z)\big),$$

where $\pi_M : U_\delta(M) \to M$ is the nearest neighbor projection onto $M = G^{-1}(\{0\})$ that is defined in a tubular neighborhood $U_\delta(M)$ of M for some $\delta > 0$ if $M \in C^2$. The step-size $\tau > 0$ has to be sufficiently small in order to guarantee the well-posedness of the iteration.

Example 4.20 (*Harmonic maps*) Minimizing the Dirichlet energy in the set

$$\mathscr{A}_h = \{u_h \in \mathscr{S}^1(\mathscr{T}_h)^m : u_h|_{\Gamma_\mathrm{D}} = u_{\mathrm{D},h},\ |u_h(z)| = 1 \quad \text{for all } z \in \mathscr{N}_h\}$$

corresponds to the situation of Theorem 4.10 with $G(s) = |s|^2 - 1$ and $M = S^{m-1} = \{s \in \mathbb{R}^m : |s| = 1\}$. In particular, we have $DG(s) = 2s$ and $\|D^2G\|_{L^\infty(\mathbb{R}^m)} = 2m^{1/2}$. The discrete tangent spaces are given by

$$T_{u_h}\mathscr{A}_h = \{w_h \in \mathscr{S}^1_\mathrm{D}(\mathscr{T}_h)^m : u_h(z) \cdot w_h(z) = 0 \quad \text{for all } z \in \mathscr{N}_h \setminus \Gamma_\mathrm{D}\}.$$

The nearest neighbor projection π_{S^2} is for $s \in \mathbb{R}^m \setminus \{0\}$ defined by $\pi_{S^2}(s) = s/|s|$.

4.3.3 Nonsmooth Equations

We consider an abstract equation of the form

$$F_h(u_h)[v_h] = 0$$

for all $v_h \in X_h$ with a continuous operator $F_h : X_h \to Y_h$ that may not be continuously differentiable. The goal is to formulate conditions that allow us to prove convergence of an appropriate generalization of the Newton method. We let X_h and Y_h be Banach spaces in the following, and assume that X_h is equipped with the norm of a Banach space X. We let $\mathrm{L}(X_h, Y_h)$ denote the space of continuous linear operators $A_h : X_h \to Y_h$ and let $\|A_h\|_{\mathrm{L}(X_h,Y_H)}$ be the corresponding operator norm.

Definition 4.3 We say that $F_h : X_h \to Y_h$ is *Newton differentiable* at $v_h \in X_h$ if there exists $\varepsilon > 0$ and a function $G_h : B_\varepsilon(v_h) \to \mathrm{L}(X_h, Y_h)$ such that

$$\lim_{w_h \to 0} \frac{\|F_h(v_h + w_h) - F_h(v_h) - G_h(v_h + w_h)[w_h]\|_{Y_h}}{\|w_h\|_X} = 0.$$

The function G_h is called the *Newton derivative* of F_h at v_h.

Remark 4.14 Notice that in contrast to the definition of the classical derivative, here the derivative is evaluated at the perturbed point $v_h + w_h$. This is precisely the expression that arises in the convergence analysis of the classical Newton iteration.

Examples 4.21 (i) If $F_h : X_h \to Y_h$ is continuously differentiable in a neighborhood of $v_h \in X_h$, then F_h is Newton differentiable at v_h with Newton derivative

$G_h = DF_h$, i.e., we have

$$\big\| F_h(v_h + w_h) - F_h(v_h) - G(v_h + w_h)[w_h] \big\|_{Y_h} \le \big\| F_h(v_h + w_h) - F_h(v_h) \\ - DF_h(v_h)[w_h] \big\|_{Y_h} + \big\| \big(DF_h(v_h) - DF(v_h + w_h)\big)[w_h] \big|_{Y_h}$$

and the right-hand side converges faster to 0 than $\|w_h\|_X$ as $w_h \to 0$.
(ii) If X_h is a Hilbert space the function $F_h(v) = \|v\|_X$, $v \in X_h$, is Newton differentiable with

$$G_h(v) = \begin{cases} v/|v| & \text{if } v \neq 0, \\ \xi & \text{if } v = 0, \end{cases}$$

where $\xi \in X_h$ with $\|\xi\|_X \le 1$ is arbitrary.
(iii) The function $F_h : \mathbb{R} \to \mathbb{R}$, $s \mapsto \max\{0, s\}$, is Newton differentiable with Newton derivative $G_h(s) = 0$ for $s < 0$, $G_h(0) = \delta$ for arbitrary $\delta \in [0, 1]$, and $G(s) = 1$ for $s > 0$.
(iv) If $1 \le p < q \le \infty$, the mapping

$$F_h : L^q(\Omega) \to L^p(\Omega), \quad v \mapsto \max\{0, v(x)\}$$

is Newton differentiable with the Newton derivative $G_h(v_h)$ for G_h as above. For $p = q$ this is false.

The semismooth Newton method is similar to the classical Newton iteration but employs the Newton derivative instead of the classical derivative.

Algorithm 4.4 (*Semismooth Newton method*) *Given $u_h^0 \in X_h$, compute the sequence $(u_h^j)_{j=0,1,\dots}$ via $u_h^{j+1} = u_h^j + d_h^j$ with $d_h^j \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m$ such that*

$$G_h(u_h^j)[d_h^j, v_h] = -F_h(u_h^j)[v_h]$$

for all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m$. Stop the iteration if $\|d_h^j\|_H \le \varepsilon_{\mathrm{stop}}$ for a norm $\|\cdot\|_H$ on $\mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m$.

Theorem 4.11 (Superlinear convergence) *Suppose that $F_h(u_h) = 0$ and $F_h : X_h \to Y_h$ is Newton differentiable at u_h, such that the linear mapping $G(\widetilde{u}_h) : X_h \to Y_h$ is invertible with $\|G_h^{-1}(\widetilde{u}_h)\|_{\mathrm{L}(Y_h, X_h)} \le M$ for every $\widetilde{u}_h \in B_\varepsilon(u_h)$ with some $\varepsilon > 0$. Then the semismooth Newton method converges superlinearly to u_h if u_h^0 is sufficiently close to u_h, i.e., for every $\eta > 0$, there exists $J \ge 0$ such that for all $j \ge J$, we have*

$$\|u_h^{j+1} - u_h\|_X \le \eta \|u_h^j - u_h\|_X.$$

Proof Noting $d_h^j = -G_h(u_h^j)^{-1} F_h(u_h^j)$, we have

$$u_h^{j+1} - u_h = u_h^j - G_h^{-1}(u_h^j) F_h(u_h^j) - u_h$$

$$
\begin{aligned}
&= u_h^j - u_h - G_h^{-1}(u_h^j)\big(F_h(u_h^j) - F_h(u_h)\big) \\
&= -G_h^{-1}(u_h^j)\big(F_h(u_h^j) - F_h(u_h) - G_h(u_h^j)(u_h^j - u_h)\big).
\end{aligned}
$$

Writing $u_h^{j+1} = u_h + w_h^{j+1}$, we have

$$
\begin{aligned}
\|w_h^{j+1}\|_X \le \|G_h^{-1}(u_h + w_h^j)\|_{\mathrm{L}(Y_h,X_h)} \| & F_h(u_h + w_h^j) - F_h(u_h) \\
& - G_h(u_h + w_h^j) w_h^j \|_{Y_h} \\
\le M\varphi(\|w_h^j\|_X) &
\end{aligned}
$$

with a function $\varphi(s)$ satisfying $\varphi(s)/s \to 0$ as $s \to 0$. If $\|w_h^0\|_X$ is sufficiently small, e.g., $\|w_h^0\|_X \le \varepsilon/(M\theta)$ with $\theta = \max_{s\in[0,1]} \varphi(s)$, then we inductively find $u_h^j \in B_\varepsilon(u_h)$ for all $j \ge 0$ and $\|w_h^j\|_X \to 0$ as $j \to \infty$. For $J \ge 0$ such that $\phi(\|w_h^j\|_X) \le (\eta/M)\|w_h^j\|_X$ for all $j \ge J$, we verify the estimate of the theorem. □

Remark 4.15 If F_h is twice continuously differentiable so that $G_h = DF_h$ is locally Lipschitz continuous and $\|DF_h^{-1}(\widetilde{u}_h)\|_{\mathrm{L}(Y_h,X_h)} \le M$, then Algorithm 4.4 coincides with the classical Newton iteration which is locally and quadratically convergent.

4.3.4 Nonsmooth, Strongly Convex Minimization

For Banach spaces X and Y, proper, convex, and lower semicontinuous functionals $G : X \to \mathbb{R} \cup \{+\infty\}$, $F : Y \to \mathbb{R} \cup \{+\infty\}$, and a bounded, linear operator $\Lambda : X \to Y$, we consider the *saddle-point problem*

$$
\inf_{u\in X} \sup_{p\in Y'} \langle \Lambda u, p\rangle - F^*(p) + G(u) = \inf_{u\in X} \sup_{p\in Y'} L(u, p).
$$

The pair (u, p) is a saddle point for L if and only if

$$
\Lambda u \in \partial F^*(p), \quad -\Lambda' p \in \partial G(u),
$$

where $\Lambda' : Y' \to X'$ denotes the formal adjoint of Λ. The related primal and dual problem consist in the minimization of the functionals

$$
I(u) = F(\Lambda u) + G(u), \quad D(p) = -F^*(p) - G^*(-\Lambda' p).
$$

We have $I(u) - D(p) \ge 0$ for all $(u, p) \in X \times Y'$ with equality if and only if (u, p) is a saddle point for L. We assume in the following that X and Y are Hilbert spaces and identify them with their duals. The descent and ascent flows $\partial_t u = -\partial_u L(u, p)$ and $\partial_t p = \partial_p L(u, p)$, respectively, motivate the following algorithm. Further details about related nonsmooth minimization problems can be found in [14].

Algorithm 4.5 (*Primal-dual iteration*) *Let $(u^0, p^0) \in X \times Y$ and set $d_t u^0 = 0$. Compute the sequences $(u^j)_{j=0,1,\ldots}$ and $(p^j)_{j=0,1,\ldots}$ by iteratively solving the equations*

$$\begin{aligned}\widetilde{u}^{j+1} &= u^j + \tau d_t u^j,\\ -d_t p^{j+1} + \Lambda\widetilde{u}^{j+1} &\in \partial F^*(p^{j+1}),\\ -d_t u^{j+1} - \Lambda' p^{j+1} &\in \partial G(u^{j+1}).\end{aligned}$$

Stop the iteration if $\|u^{j+1} - u^j\|_X \le \varepsilon_{\text{stop}}$.

Remark 4.16 The equations in Algorithm 4.5 are equivalent to the variational inequalities

$$\begin{aligned}\langle -d_t u^{j+1} - \Lambda' p^{j+1}, v - u^{j+1}\rangle_X &\le G(v) - G(u^{j+1}) - \frac{\alpha}{2}\|v - u^{j+1}\|_X^2,\\ \langle -d_t p^{j+1} + \Lambda\widetilde{u}^{j+1}, q - p^{j+1}\rangle_Y &\le F^*(q) - F^*(p^{j+1})\end{aligned}$$

for all $(v, q) \in X \times Y$. Here, $\alpha > 0$ if G is uniformly convex.

We prove convergence of Algorithm 4.5 assuming that $\alpha > 0$. We abbreviate by $\|\Lambda\|$ the operator norm $\|\Lambda\|_{\mathrm{L}(X,Y)}$.

Theorem 4.12 (Convergence) *Let (u, p) be a saddle point for L. If $\tau\|\Lambda\| \le 1$, we have for every $J \ge 0$ that*

$$\frac{1-\tau\|\Lambda\|}{2}\|p - p^{J+1}\|_Y^2 + \frac{1}{2}\|u - u^{J+1}\|_X^2 + \tau\sum_{j=0}^{J}\frac{\alpha}{2}\|u - u^{j+1}\|_X^2 \le \frac{1}{2}\|p - p^0\|_Y^2 + \frac{1}{2}\|u - u^0\|_X^2.$$

In particular, the iteration of Algorithm 4.5 terminates.

Proof We denote $\delta_u^{j+1} = u - u^{j+1}$ and $\delta_p^{j+1} = p - p^{j+1}$ in the following. Using that $d_t\delta_u^{j+1} = -d_t u^{j+1}$ and $d_t\delta_p^{j+1} = -d_t p^{j+1}$, we find that

$$\begin{aligned}\Upsilon(j+1) &= \frac{d_t}{2}\big(\|\delta_p^{j+1}\|_Y^2 + \|\delta_u^{j+1}\|_X^2\big) + \frac{\tau}{2}\big(\|d_t\delta_u^{j+1}\|_X^2 + \|d_t\delta_p^{j+1}\|_Y^2\big) + \frac{\alpha}{2}\|\delta_u^{j+1}\|_X^2\\ &= \langle d_t\delta_p^{j+1}, \delta_p^{j+1}\rangle_Y + \langle d_t\delta_u^{j+1}, \delta_u^{j+1}\rangle_X + \frac{\alpha}{2}\|\delta_u^{j+1}\|_X^2\\ &= -\langle d_t p^{j+1}, p - p^{j+1}\rangle_Y - \langle d_t u^{j+1}, u - u^{j+1}\rangle_X + \frac{\alpha}{2}\|u - u^{j+1}\|_X^2.\end{aligned}$$

The equations for $d_t p^{j+1}$ and $d_t u^{j+1}$ of Algorithm 4.5 and their equivalent characterization in Remark 4.16 lead to

$$\begin{aligned}\Upsilon(j+1) \le F^*(p) - F^*(p^{j+1}) - \langle \Lambda \widetilde{u}^{j+1}, p - p^{j+1}\rangle_Y \\ + G(u) - G(u^{j+1}) + \langle \Lambda' p^{j+1}, u - u^{j+1}\rangle_X \\ = \big[\langle \Lambda u, p^{j+1}\rangle_Y - F^*(p^{j+1}) + G(u)\big] \\ - \big[\langle \Lambda u^{j+1}, p\rangle_Y - F^*(p) + G(u^{j+1})\big] + \langle \Lambda u^{j+1}, p\rangle_Y \\ - \langle \Lambda \widetilde{u}^{j+1}, p - p^{j+1}\rangle_Y - \langle \Lambda' p^{j+1}, u^{j+1}\rangle_Y .\end{aligned}$$

The definitions of $F^{**} = F$ and G^* imply that

$$\begin{aligned}\langle \Lambda u, p^{j+1}\rangle_Y - F^*(p^{j+1}) \le F(\Lambda u), \\ -\langle u^{j+1}, \Lambda' p\rangle_X - G(u^{j+1}) \le G^*(-\Lambda' p).\end{aligned}$$

These estimates and the identity $u^{j+1} - \widetilde{u}^{j+1} = \tau^2 d_t^2 u^{j+1} = -\tau^2 d_t^2 \delta_u^{j+1}$ allow us to deduce that

$$\begin{aligned}\Upsilon(j+1) \le F(\Lambda u) + G(u) + F^*(p) + G^*(-\Lambda' p) \\ + \langle \Lambda u^{j+1}, p\rangle_Y - \langle \Lambda \widetilde{u}^{j+1}, p - p^{j+1}\rangle_Y - \langle \Lambda' p^{j+1}, u^{j+1}\rangle_X \\ = I(u) - D(p) - \tau^2 \langle \Lambda d_t^2 \delta_u^{j+1}, \delta_p^{j+1}\rangle_Y .\end{aligned}$$

We use $I(u) - D(p) = 0$ to derive the estimate

$$\Upsilon(j+1) \le -\tau^2 \langle \Lambda d_t^2 \delta_u^{j+1}, \delta_p^{j+1}\rangle_Y .$$

A summation of the estimate over $j = 0, 1, \dots, J$ and multiplication by τ lead to

$$\begin{aligned}\frac{1}{2}\big(\|\delta_p^{J+1}\|_Y^2 + \|\delta_u^{J+1}\|_X^2\big) + \frac{\tau^2}{2}\sum_{j=0}^{J}\big(\|d_t \delta_u^{j+1}\|_X^2 + \|d_t \delta_p^{j+1}\|_Y^2\big) + \frac{\alpha}{2}\sum_{j=0}^{J}\|\delta_u^{j+1}\|_X^2 \\ \le \frac{1}{2}\big(\|\delta_p^0\|_Y^2 + \|\delta_u^0\|_X^2\big) - \tau^3 \sum_{j=0}^{J}\langle \Lambda d_t^2 \delta_u^{j+1}, \delta_p^{j+1}\rangle .\end{aligned}$$

A summation by parts, $-d_t \delta_u^0 = d_t u^0 = 0$, and Young's inequality show that

$$\begin{aligned}&- \tau^3 \sum_{j=0}^{J}\langle \Lambda d_t^2 \delta_u^{j+1}, \delta_p^{j+1}\rangle_Y = \tau^3 \sum_{j=0}^{J}\langle \Lambda d_t \delta_u^{j}, d_t \delta_p^{j+1}\rangle_Y + \tau^2 \langle \Lambda d_t \delta_u^j, \delta_p^j\rangle_Y \big|_{j=0}^{J+1} \\ &\le \frac{\tau^2}{2}\Big(\sum_{j=0}^{J}\tau^2 \|\Lambda d_t \delta_u^j\|_Y^2 + \|d_t \delta_p^{j+1}\|_Y^2\Big) + \frac{\tau\|\Lambda\|}{2}\|\delta_p^{J+1}\|_Y^2 + \frac{\tau^3}{2\|\Lambda\|}\|\Lambda d_t \delta_u^{J+1}\|_Y^2 .\end{aligned}$$

A combination of the estimates proves the theorem. □

Remarks 4.17 (i) The assumption that a saddle point exists implies that primal and dual problem are related by a strong duality principle.
(ii) If F is strongly convex and G is only convex, then the roles of u and p have to be exchanged to ensure convergence.
(iii) The algorithm may be regarded as an inexact Uzawa algorithm. The classical Uzawa method corresponds to omitting $d_t u^{j+1}$, i.e., solving the equation $-\Lambda' p^{j+1} \in \partial G(u^{j+1})$ for u^{j+1} at every step of the algorithm.
(iv) Algorithm 4.5 is practical if the proximity operators $r = (1 + \tau \partial F^*)^{-1} q$ and $w = (1 + \tau \partial G)^{-1} v$ can be easily evaluated, i.e., if the unique minimizers of

$$w \mapsto \frac{1}{2\tau} \|w - v\|_X^2 + G(w), \quad r \mapsto \frac{1}{2\tau} \|r - q\|_Y^2 + F^*(r),$$

are directly accessible. This is the case for quadratic functionals and indicator functionals.

Example 4.22 In the case of the discretized Poisson problem with $X = \mathscr{S}_0^1(\mathscr{T}_h)$, we may choose $Y = \mathscr{L}^0(\mathscr{T}_h)^d$, $\Lambda = \nabla$,

$$F(p_h) = \frac{1}{2} \int_\Omega |p_h|^2 \, \mathrm{d}x, \quad G(u_h) = \int_\Omega f u_h \, \mathrm{d}x,$$

and exchange the roles of u_h and p_h. Letting $P_{h,0} f$ denote the L^2 projection onto $\mathscr{S}_0^1(\mathscr{T}_h)$, the iteration reads

$$\begin{aligned} \widetilde{p}_h^{j+1} &= p_h^j + \tau d_t p_h^{j+1}, \\ -d_t u_h^{j+1} + \operatorname{div}_h \widetilde{p}_h^{j+1} &= P_h f, \\ -d_t p_h^{j+1} + \nabla u_h^{j+1} &= p_h^{j+1}. \end{aligned}$$

The discrete divergence operator $\operatorname{div}_h : \mathscr{L}^0(\mathscr{T}_h)^d \to \mathscr{S}_0^1(\mathscr{T}_h)$ is for every elementwise constant vector field $q_h \in \mathscr{L}^0(\mathscr{T}_h)^d$ defined by $(\operatorname{div}_h q_h, v_h) = -(q_h, \nabla v_h)$ for all $v_h \in \mathscr{S}_0^1(\mathscr{T}_h)$. Convergence holds if $\tau \|\nabla\| \le 1$, where $\|\nabla\| \le c h^{-1}$.

4.3.5 Nested Iteration

The semismooth and classical Newton method can only be expected to converge if the starting value u_h^0 is sufficiently close to the discrete solution u_h. The radius of the ball around u_h which contains such starting values may depend critically on the mesh-size in the sense that it becomes smaller when the mesh is refined. Such a behavior reflects the problem that the Newton scheme may not be well-defined for the underlying continuous formulation. When a sequence of refined triangulations is used, the corresponding finite element spaces are nested, and one may use an approximate solution computed on a coarse grid to define a starting value for the iteration process on the finer grid. Besides providing a method to construct feasible starting values, this approach can also significantly reduce the computational effort.

Algorithm 4.6 (*Nested iteration*) *Let* $(\mathscr{T}_\ell)_{\ell=0,\dots,L}$ *be a sequence of triangulations with* $\mathscr{S}^1(\mathscr{T}_{\ell-1}) \subset \mathscr{S}^1(\mathscr{T}_\ell)$ *for* $\ell = 1, 2, \dots, L$. *Set* $\ell = 0$ *and choose* $u_\ell^0 \in \mathscr{S}^1(\mathscr{T}_\ell)$.
(i) Iteratively approximate a solution $u_\ell \in \mathscr{S}^1(\mathscr{T}_\ell)$ *of* $F_\ell(u_\ell) = 0$ *using the starting value* u_ℓ^0 *to obtain an approximate solution* $u_\ell^* \in \mathscr{S}^1(\mathscr{T}_\ell)$.
(ii) Stop if $\ell = L$. *Otherwise set* $u_{\ell+1}^0 = u_\ell^*$, $\ell \to \ell + 1$, *and continue with* (i).

We make the ideas more precise for a red-green-blue refinement method. The definition is easily generalized to other refinement methods such as newest-vertex bisection.

Definition 4.4 We say that $\mathscr{T}_h$ is a *refinement* of the triangulation $\mathscr{T}_H$ if $\mathscr{S}^1(\mathscr{T}_H) \subset \mathscr{S}^1(\mathscr{T}_h)$ and for every node $z^h \in \mathscr{N}_h$ we either have $z^h \in \mathscr{N}_H$ or there exist nodes $z_1^H, z_2^H \in \mathscr{N}_H$ with $z^h = (z_1^H + z_2^H)/2$, cf. Fig. 4.3.

Lemma 4.1 (Prolongation) *Let* $\mathscr{T}_h$ *be a refinement of the triangulation* $\mathscr{T}_H$. *Given* $u_H \in \mathscr{S}^1(\mathscr{T}_H)$, *we have* $u_h = u_H \in \mathscr{S}^1(\mathscr{T}_h)$ *with nodal values* $u_h(z^h) = u_H(z^h)$ *for every* $z^h \in \mathscr{N}_H \subset \mathscr{N}_h$ *and* $u_h(z^h) = (u_H(z_1^H) + u_H(z_2^H))/2$ *for every* $z^h \in \mathscr{N}_h \setminus \mathscr{N}_H$ *and* $z_1^h, z_2^h \in \mathscr{N}_H$ *with* $z^h = (z_1^H + z_2^H)/2$. *In particular, there exists a linear* prolongation operator

$$Pr_{H\to h}^1 : \mathbb{R}^{\mathscr{N}_H} \to \mathbb{R}^{\mathscr{N}_h}, \quad \big(u_H(z^H)\big)_{z^H \in \mathscr{N}_H} \mapsto \big(u_H(z^h)\big)_{z^h \in \mathscr{N}_h}$$

for every $u_H \in \mathscr{S}^1(\mathscr{T}_H)$.

Proof The assertion of the lemma follows from the fact that the function u_h is affine on every one-dimensional subsimplex in the triangulation. □

Remarks 4.18 (i) The superscript 1 in $Pr_{H\to h}^1$ corresponds to affine functions. Analogously, there exists a linear operator $Pr_{H\to h}^0$ that maps the values of an elementwise constant function on $\mathscr{T}_H$ to the values of the function represented on $\mathscr{T}_h$.
(ii) Matrices that realize the linear mappings of the nodal or elementwise values are provided by the routine `red_refine.m`.
(iii) Nested iterations are the simplest version of a multigrid scheme. In more general versions, grid transfer from a fine to a coarse grid called restriction is required. This is often realized with the adjoint operators, i.e., with the transposed matrices.
(iv) For nonnested finite element spaces the grid transfer can be realized with interpolation or projection operators.

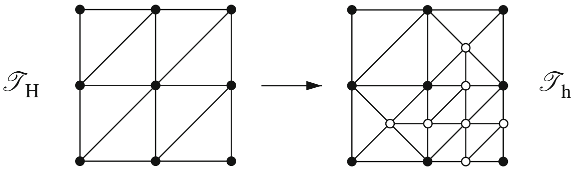


Fig. 4.3 The nodes of the refined triangulation are either existing nodes (*dots*) or midpoints of bisected edges (*circles*)

References

1. Babuška, I.: Error-bounds for finite element method. Numer. Math. **16**, 322–333 (1970/1971)
2. Berger, M.S.: Nonlinearity and Functional Analysis. Academic Press, New York (1977)
3. Boffi, D., Brezzi, F., Fortin, M.: Mixed Finite Element Methods and Applications. Springer Series in Computational Mathematics, vol. 44. Springer, Heidelberg (2013)
4. Bonnans, J.F., Gilbert, J.C., Lemaréchal, C., Sagastizábal, C.A.: Numerical Optimization, 2nd edn. Universitext. Springer, Berlin (2006)
5. Braides, A.: Approximation of Free-Discontinuity Problems. Lecture Notes in Mathematics, vol. 1694. Springer, Berlin (1998)
6. Brenner, S.C.: A Cautionary Tale in Numerical PDEs. In: Sonia Kovalevskii Lecture, ICIAM 2011. Vancouver, Canada (2011)
7. Brezzi, F.: On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge **8**(R-2), 129–151 (1974)
8. Costabel, M.: A coercive bilinear form for Maxwell's equations. J. Math. Anal. Appl. **157**(2), 527–541 (1991). http://dx.doi.org/10.1016/0022-247X(91)90104-8
9. Dacorogna, B.: Direct Methods in the Calculus of Variations. Applied Mathematical Sciences, vol. 78, 2nd edn. Springer, New York (2008)
10. Dziuk, G., Hutchinson, J.E.: Finite element approximations to surfaces of prescribed variable mean curvature. Numer. Math. **102**(4), 611–648 (2006). http://dx.doi.org/10.1007/s00211-005-0649-7
11. Evans, L.C.: Weak Convergence Methods for Nonlinear Partial Differential Equations. CBMS Regional Conference Series in Mathematics, vol. 74. Published for the Conference Board of the Mathematical Sciences, Washington (1990)
12. Ito, K., Kunisch, K.: Lagrange Multiplier Approach to Variational Problems and Applications. Advances in Design and Control, vol. 15. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2008)
13. Nečas, J.: Sur une méthode pour résoudre les équations aux dérivées partielles du type elliptique, voisine de la variationnelle. Ann. Scuola Norm. Sup. Pisa **16**(3), 305–326 (1962)
14. Nesterov, Y., Nemirovski, A.: On first-order algorithms for ℓ_1/nuclear norm minimization. Acta Numer. **22**, 509–575 (2013)
15. Nochetto, R.H., Savaré, G., Verdi, C.: A posteriori error estimates for variable time-step discretizations of nonlinear evolution equations. Commun. Pure Appl. Math. **53**(5), 525–589 (2000)
16. Repin, S.: A Posteriori Estimates for Partial Differential Equations. Radon Series on Computational and Applied Mathematics, vol. 4. Walter de Gruyter GmbH & Co. KG, Berlin (2008)
17. Rulla, J.: Error analysis for implicit approximations to solutions to Cauchy problems. SIAM J. Numer. Anal. **33**(1), 68–87 (1996). http://dx.doi.org/10.1137/0733005
18. Struwe, M.: Variational Methods, 4th edn. Springer, Berlin (2008)

Part II
Approximation of Classical Formulations

Chapter 5
The Obstacle Problem

5.1 Analytical Properties

We discuss in this section analytical properties of the *obstacle problem* which is the prototypical example of a convex minimization problem with an inequality constraint that leads to a variational inequality. Throughout this chapter, for $f \in L^2(\Omega)$, $g \in L^2(\Gamma_{\rm N})$, and $\chi \in H^1(\Omega)$ with $\chi \le 0$ on $\Gamma_{\rm D}$, we consider the problem of minimizing the functional

$$I(u) = \frac{1}{2}\int_\Omega |\nabla u|^2 \,{\rm d}x - \int_\Omega f u \,{\rm d}x - \int_{\Gamma_{\rm N}} g u \,{\rm d}s$$

in the set of functions $u \in K$ defined by the convex set

$$K = \{v \in H^1_{\rm D}(\Omega) : v \ge \chi \text{ almost everywhere in } \Omega\}.$$

The model problem is illustrated in Fig. 5.1. We show that minimizers for this constrained minimization problem are unique and have certain regularity properties. For more general statements and other minimization problems with inequality constraints, we refer the reader to the textbooks [4, 6, 7].

In addition to the general assumption regarding homogeneous Dirichlet conditions on $\Gamma_{\rm D}$, justified by the splitting $u = \widetilde{u} + \widetilde{u}_{\rm D}$ with the unknown $\widetilde{u} \in H^1_{\rm D}(\Omega)$ and the extension $\widetilde{u}_{\rm D} \in H^1(\Omega)$ of $u_{\rm D}$, we will often consider the case $g = 0$ on $\Gamma_{\rm N}$. This is justified by requiring that $\partial_n \widetilde{u}_{\rm D} = g$ on $\Gamma_{\rm N}$.

5.1.1 *Existence and Uniqueness*

We apply the direct method in the calculus of variations to establish the existence of a solution. The main ingredients for the proof are the strict convexity and continuity of I that imply the weak lower semicontinuity and weak closedness of K.

S. Bartels, *Numerical Methods for Nonlinear Partial Differential Equations*,
Springer Series in Computational Mathematics 47,
DOI 10.1007/978-3-319-13797-1_5

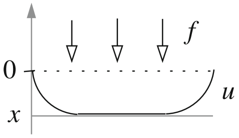


Fig. 5.1 Deflection u of a membrane due to a force f and constrained by a constant obstacle χ

Theorem 5.1 (Well posedness) *There exists a unique minimizer $u \in K$ for I.*

Proof The functional I is weakly lower semicontinuous on $H^1_{\mathrm{D}}(\Omega)$. Its boundedness from below follows from Hölder's inequality, i.e.,

$$I(u) \geq \frac{1}{2}\|\nabla u\|^2 - \|f\|\|u\| - \|g\|_{L^2(\Gamma_{\mathrm{N}})}\|u\|_{L^2(\Gamma_{\mathrm{N}})}.$$

With the Poincaré inequality $\|u\| \leq c_P\|\nabla u\|$, the trace inequality $\|u\|_{L^2(\Gamma_{\mathrm{N}})} \leq c_T\|\nabla u\|$, and Young's inequality, we deduce that

$$\begin{aligned} I(u) &\geq \frac{1}{2}\|\nabla u\|^2 - 2c_P^2\|f\|^2 - \frac{1}{8}\|\nabla u\|^2 - 2c_T^2\|g\|^2_{L^2(\Gamma_{\mathrm{N}})} - \frac{1}{8}\|\nabla u\|^2 \\ &\geq \frac{1}{4}\|\nabla u\|^2 - 2c_P^2\|f\|^2 - 2c_T^2\|g\|^2_{L^2(\Gamma_{\mathrm{N}})}. \end{aligned}$$

This proves that I is coercive and bounded from below on $H^1_{\mathrm{D}}(\Omega)$. Since the function $\widetilde{u} = \max\{0, \chi\}$ satisfies $\widetilde{u} \in K$ there exists an infimizing sequence $(u_j)_{j\in\mathbb{N}} \subset K$ with $\lim_{j\to\infty} I(u_j) = \inf_{v\in K} I(v)$. The coercivity of I shows that this sequence is bounded and hence there exists a weakly convergent subsequence $(u_{j_k})_{k\in\mathbb{N}} \subset K$ and a weak limit $u \in H^1_{\mathrm{D}}(\Omega)$. To show that $u \in K$ we notice that by the compact embedding of $H^1_{\mathrm{D}}(\Omega)$ into $L^2(\Omega)$ we have $u_{j_k} \to u$ in $L^2(\Omega)$, and for a further subsequence that $u_{j_{k_\ell}}(x) \to u(x)$ for almost every $x \in \Omega$. Since $u_j(x) \geq \chi(x)$ for every $j \in \mathbb{N}$ and almost every $x \in \Omega$ we conclude that $u \geq \chi$ almost everywhere in Ω. The weak lower semicontinuity of I shows that

$$I(u) \leq \liminf_{k\to\infty} I(u_{j_k}) = \lim_{j\to\infty} I(u_j) = \inf_{v\in K} I(v).$$

Thus, $u \in K$ is a solution for the minimization problem. To show that the minimizer is unique, we let $u_1, u_2 \in K$ be minimizers and notice that by the convexity of K, we have $(u_1 + u_2)/2 \in K$ and

$$\frac{1}{2}I(u_1) + \frac{1}{2}I(u_2) - I\big(\frac{u_1 + u_2}{2}\big) = \frac{1}{8}\|\nabla(u_1 - u_2)\|^2.$$

If $u_1 \neq u_2$, then the right-hand side is positive which leads to the contradiction $I\big((u_1 + u_2)/2\big) < \big(I(u_1) + I(u_2)\big)/2$. Therefore, I has a unique minimizer. □

5.1.2 Equivalent Formulations

We want to formulate optimality conditions for a minimizer $u \in K$ of the obstacle problem. Due to the convexity of K, we have for every $v \in K$ and $t \in [0, 1]$ that $u + t(v - u) \in K$ with

$$I(u) \le I\big(u + t(v - u)\big).$$

Setting $\phi(t) = I\big(u + t(v - u)\big)$ we have that ϕ is increasing on [0, 1] and thus its right-sided derivative at 0 is nonnegative. More precisely, for $t > 0$ we have

$$\begin{aligned}
0 &\le t^{-1}\big(I\big(u + t(v - u)\big) - I(u)\big)\\
&= \frac{1}{2t}\int_\Omega |\nabla(u + t(v - u))|^2 \,\mathrm{d}x - \frac{1}{2t}\int_\Omega |\nabla u|^2 \,\mathrm{d}x\\
&\quad - \int_\Omega f(v - u)\,\mathrm{d}x - \int_{\Gamma_\mathrm{N}} g(v - u)\,\mathrm{d}s\\
&= \int_\Omega \nabla u \cdot \nabla(v - u)\,\mathrm{d}x + \frac{t}{2}\int_\Omega |\nabla(v - u)|^2 \,\mathrm{d}x\\
&\quad - \int_\Omega f(v - u)\,\mathrm{d}x - \int_{\Gamma_\mathrm{N}} g(v - u)\,\mathrm{d}s.
\end{aligned}$$

Considering the limit $t \to 0$ we find that the *variational inequality*

$$\int_\Omega \nabla u \cdot \nabla(v - u)\,\mathrm{d}x \ge \int_\Omega f(v - u)\,\mathrm{d}x + \int_{\Gamma_\mathrm{N}} g(v - u)\,\mathrm{d}s$$

holds for all $v \in K$. The arguments also show that this formulation is a sufficient characterization of a minimizer. If the variational inequality is satisfied, then we have for every $v \in K$ that

$$I(u) - I(v) \le \int_\Omega \nabla u \cdot \nabla(u - v)\,\mathrm{d}x - \int_\Omega f(u - v)\,\mathrm{d}x - \int_{\Gamma_\mathrm{N}} g(u - v)\,\mathrm{d}s \le 0,$$

i.e., $u \in K$ is minimal for I. An alternative characterization of the solution employs the *indicator functional* $I_K : H^1_\mathrm{D}(\Omega) \to \mathbb{R} \cup \{+\infty\}$ defined by

$$I_K(v) = \begin{cases} 0 & \text{for } v \in K,\\ +\infty & \text{for } v \notin K.\end{cases}$$

We then include the constraint $u \in K$ in the minimization problem by considering the functional

$$\widetilde{I}(u) = I(u) + I_K(u)$$

on $H^1_{\mathrm{D}}(\Omega)$. A minimizer $u \in H^1_{\mathrm{D}}(\Omega)$ for $\widetilde{I}$ satisfies $u \in K$ and is a minimizer for I in K. One verifies directly that $\widetilde{I}$ is a convex functional which is weakly lower semicontinuous. With

$$\partial\widetilde{I}(u) = \{\mu \in H^1_{\mathrm{D}}(\Omega)' : \langle \mu, v - u\rangle \le \widetilde{I}(v) - \widetilde{I}(u) \text{ for all } v \in H^1_{\mathrm{D}}(\Omega)\}$$

we have that $u \in K$ is a minimizer for $\widetilde{I}$ if and only if $0 \in \partial\widetilde{I}(u)$. Since I is Fréchet differentiable we have

$$0 \in DI(u) + \partial I_K(u).$$

Equivalently, there exists a *Lagrange multiplier* $\lambda \in \partial I_K(u)$ such that $0 = DI(u)[w] + \langle\lambda, w\rangle$ for all $w \in H^1_{\mathrm{D}}(\Omega)$. This means that

$$0 = \int_\Omega \nabla u \cdot \nabla w \,\mathrm{d}x - \int_\Omega f w \,\mathrm{d}x - \int_{\Gamma_{\mathrm{N}}} g w \,\mathrm{d}s + \langle\lambda, w\rangle$$

for all $w \in H^1_{\mathrm{D}}(\Omega)$. With $v = u + \phi \in K$ for every $\phi \in H^1_{\mathrm{D}}(\Omega)$ with $\phi \ge 0$ in Ω, we deduce from the variational inequality above that

$$\langle\lambda, \phi\rangle = -\int_\Omega \nabla u \cdot \nabla\phi \,\mathrm{d}x + \int_\Omega f\phi \,\mathrm{d}x + \int_{\Gamma_{\mathrm{N}}} g\phi \,\mathrm{d}s \le 0,$$

i.e., that $\lambda \le 0$ in the distributional sense. The variational inequality is an equality in the set $\{x \in \Omega : u(x) > \chi(x)\}$ and therefore $\operatorname{supp}\lambda \subset \{x \in \Omega : u(x) = \chi(x)\}$. We summarize the observations in the following theorem.

Theorem 5.2 (Variational inequality) *A function $u \in K$ is the unique minimizer for I for functions in K if and only if*

$$\int_\Omega \nabla u \cdot \nabla(v - u)\,\mathrm{d}x \ge \int_\Omega f(v - u)\,\mathrm{d}x + \int_{\Gamma_{\mathrm{N}}} g(v - u)\,\mathrm{d}s$$

for all $v \in K$. This is satisfied if and only if there exists $\lambda \in H^1_{\mathrm{D}}(\Omega)'$ with $\operatorname{supp}\lambda \subset \{x \in \Omega : u(x) = \chi(x)\}$ *and $\lambda \le 0$ such that*

$$\int_\Omega \nabla u \cdot \nabla w \,\mathrm{d}x + \langle\lambda, w\rangle = \int_\Omega f w \,\mathrm{d}x + \int_{\Gamma_{\mathrm{N}}} g w \,\mathrm{d}s$$

for all $w \in H^1_{\mathrm{D}}(\Omega)$.

Proof The equivalence of the variational inequality and the minimization problem has been discussed above. That the variational inequality implies the equation with the Lagrange multiplier follows by defining λ as above. The converse implication is a consequence of the fact that $v - u \geq 0$ in the set $\{x \in \Omega : u(x) = \chi(x)\}$ for every $v \in K$. □

The region in which the solution u is in contact with the obstacle χ is of importance in many applications.

Definition 5.1 For the solution of the obstacle problem we define the *contact zone* or *coincidence set* by

$$\mathscr{C} = \{x \in \Omega : u(x) = \chi(x)\}.$$

The boundary $\partial\mathscr{C} \cap \Omega$ is called the *free boundary*.

Remarks 5.1 (i) By choosing $\lambda \in H^1_{\mathrm{D}}(\Omega)$ such that $(\nabla\lambda, \nabla w) = \langle\lambda, w\rangle$ for all $w \in H^1_{\mathrm{D}}(\Omega)$, and setting $\widetilde{\lambda} = -\Delta\lambda$ and $\widetilde{\lambda}_{\mathrm{N}} = \partial_n\lambda$, we find that the strong form of the obstacle problem reads

$$-\Delta u + \widetilde{\lambda} = f, \;\; u \geq \chi, \;\; \widetilde{\lambda}, \, \widetilde{\lambda}_{\mathrm{N}} \leq 0, \;\; \operatorname{supp}\widetilde{\lambda} \subset \mathscr{C}, \;\; \partial_n u|_{\Gamma_{\mathrm{N}}} + \widetilde{\lambda}_{\mathrm{N}}|_{\Gamma_{\mathrm{N}}} = g, \;\; u|_{\Gamma_{\mathrm{D}}} = 0.$$

(ii) In the complement of the contact zone we have $-\Delta u = f$ and the Lagrange multiplier is supported in $\mathscr{C}$ in the sense that $\langle\lambda, w\rangle = 0$ if $\operatorname{supp} w \subset \mathscr{C}^c$.
(iii) We have the *complementarity principle*

$$(\Delta u + f)(u - \chi) = 0,$$

i.e., we have $u = \chi$ or $-\Delta u = f$ almost everywhere in Ω.

5.1.3 Regularity

It is not obvious that solutions of the obstacle problem obey higher regularity properties. In one-dimensional situations, continuity of the derivative of the solution, i.e., $u \in H^2(\Omega)$, can be verified directly.

Proposition 5.1 (One-dimensional regularity) *Let $\Omega = (a, b) \subset \mathbb{R}$, $\chi \in H^2(a, b)$, $f \in L^2(a, b)$, and $u \in H^1_0(a, b)$ be such that $u \geq \chi$ in (a, b) and*

$$\int_a^b u'(v - u)' \, \mathrm{d}x \geq \int_a^b f(v - u) \, \mathrm{d}x$$

for all $v \in H^1_0(a, b)$ with $v \geq \chi$ in (a, b). Then $u \in H^2(a, b)$.

Proof We have that $u = \chi$ in $\mathscr{C}$ and $-u'' = f$ in $\mathscr{C}^c$. It thus suffices to consider a point $x_c \in \partial\mathscr{C}$ and show that u' is continuous at x_c. By definition there exists $\varepsilon > 0$ such that $u(x_c - y) = \chi(x_c - y)$ and $u(x_c + y) > \chi(x_c + y)$ either for all $-\varepsilon < y \le 0$ or for all $0 \le y < \varepsilon$. Without loss of generality we consider the latter situation. For every nonnegative function $\phi \in C^1(\Omega)$ with $\operatorname{supp}\phi \subset B_\varepsilon(x_c)$, we have $u + \phi \ge \chi$ and thus

$$\begin{aligned} 0 &\le \int\limits_{x_c-\varepsilon}^{x_c+\varepsilon} u'\phi' \,\mathrm{d}x - \int\limits_{x_c-\varepsilon}^{x_c+\varepsilon} f\phi \,\mathrm{d}x \\ &= -\int\limits_{x_c-\varepsilon}^{x_c} \chi''\phi \,\mathrm{d}x - \int\limits_{x_c-\varepsilon}^{x_c} f\phi \,\mathrm{d}x + \big(\chi'(x_c) - u'(x_c)\big)\phi(x_c). \end{aligned}$$

Since $\phi \ge 0$ is arbitrary, we find that $u'(x_c) \le \chi'(x_c)$ by taking the limit $\varepsilon \to 0$. The nonnegative, continuous function $\delta = u - \chi$ satisfies $\delta(x_c) = 0$ and $\delta'(x_c) \le 0$. Thus, we have $\delta'(x_c) = 0$, i.e., $u'(x_c) = \chi'(x_c)$ which implies that u' is continuous at x_c. □

Example 5.1 For $\Omega = (-1, 1)$, $\chi(x) = 1 - 4x^2/3$, it follows that the minimizer $u \in H^1_0(\Omega)$ for $I(u) = \int_{-1}^{1} |u'(x)|^2 \,\mathrm{d}x$ with $u \ge \chi$ is given by

$$u(x) = \begin{cases} 1 - 4x^2/3 & \text{for } |x| \le 1/2, \\ (4/3)(1 - |x|) & \text{for } |x| \ge 1/2, \end{cases}$$

cf. Fig. 5.2. In particular, $\mathscr{C} = [-1/2, 1/2]$. We note that $u \in H^2(\Omega)\setminus H^3(\Omega)$.

A similar result holds in higher-dimensional situations.

Theorem 5.3 (Regularity [1]) *If $\Gamma_{\mathrm{D}} = \partial\Omega$, $\chi \in H^2(\Omega)$, $\chi|_{\partial\Omega} \le 0$, and Ω is convex, then $u \in H^2(\Omega)$.*

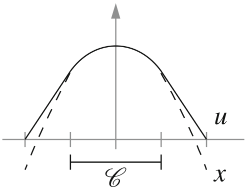


Fig. 5.2 Solution u of a one-dimensional obstacle problem with obstacle χ; the slopes of the obstacle and the displacement coincide at the boundary of the contact zone $\mathscr{C}$

5.1.4 Penalization

Penalty methods provide an attractive way to include constraints in a minimization problem and approximate the original formulation by a sequence of continuous, Fréchet-differentiable functionals. In particular, standard algorithms like descent methods can be used for an approximate solution of the functionals. The main idea is to penalize a constraint violation, e.g., via

$$I_\varepsilon(u) = \frac{1}{2}\int_\Omega |\nabla u|^2 \,\mathrm{d}x - \int_\Omega f u \,\mathrm{d}x - \int_{\Gamma_\mathrm{N}} g u \,\mathrm{d}s + \frac{\varepsilon^{-2}}{2}\int_\Omega (u-\chi)_-^2 \,\mathrm{d}x,$$

where $\varepsilon > 0$ is a small *penalty parameter* and $(s)_- = \min\{s, 0\}$ for $s \in \mathbb{R}$. A violation of the constraint $u - \chi \ge 0$ thus leads to a large contribution to the energy. In general the minimizer $u_\varepsilon \in H^1_\mathrm{D}(\Omega)$ will not satisfy $u_\varepsilon \ge \chi$, so that a *penetration* of the obstacle occurs. We include an estimate for the difference $u - u_\varepsilon$ and the penetration error for a simplified situation.

Theorem 5.4 (Nonconforming penalization) *Let $\Gamma_\mathrm{D} = \partial\Omega$ and $\chi = 0$ and assume that the solution $u \in K$ of the obstacle problem satisfies $u \in H^2(\Omega)$. For the unique minimizer $u_\varepsilon \in H^1_0(\Omega)$ of the penalized functional I_ε, we have*

$$2\|\nabla(u - u_\varepsilon)\|^2 + \varepsilon^{-2}\|u_\varepsilon^-\|^2 \le \varepsilon^2\|f + \Delta u\|^2,$$

where $u_\varepsilon^-(x) = \min\{u_\varepsilon(x), 0\}$ for almost every $x \in \Omega$.

Proof The existence of a unique minimizer $u_\varepsilon \in H^1_0(\Omega)$ follows from the direct method in the calculus of variations and the strict convexity of I_ε. The minimizer satisfies

$$(\nabla u_\varepsilon, \nabla w) + \varepsilon^{-2}(u_\varepsilon^-, w) = (f, w)$$

for all $w \in H^1_0(\Omega)$. We set $u_\varepsilon^+ = u_\varepsilon - u_\varepsilon^-$ and note that $u_\varepsilon^+ \ge 0$ in Ω. With the variational inequality satisfied by u and the equation satisfied by u_ε, we find that

$$\begin{aligned}\|\nabla(u - u_\varepsilon)\|^2 &= (\nabla u, \nabla[u - u_\varepsilon^+]) + (\nabla u, \nabla[u_\varepsilon^+ - u_\varepsilon]) - (\nabla u_\varepsilon, \nabla[u - u_\varepsilon]) \\ &\le (f, u - u_\varepsilon^+) + (\nabla u, \nabla[u_\varepsilon^+ - u_\varepsilon]) - (f, u - u_\varepsilon) \\ &\quad + \varepsilon^{-2}(u_\varepsilon^-, u - u_\varepsilon) \\ &= (f, u_\varepsilon^-) - (\nabla u, \nabla u_\varepsilon^-) + \varepsilon^{-2}(u_\varepsilon^-, u - u_\varepsilon).\end{aligned}$$

We note that $(u_\varepsilon^-, u_\varepsilon^+) = 0$ and $(u_\varepsilon^-, u) \le 0$ since $u \ge 0$ so that

$$\varepsilon^{-2}(u_\varepsilon^-, u - u_\varepsilon) \le -\varepsilon^{-2}\|u_\varepsilon^-\|^2.$$

Green's identity and Hölder and Young inequalities lead to

$$\begin{aligned}\|\nabla(u-u_\varepsilon)\|^2+\varepsilon^{-2}\|u_\varepsilon^-\|^2 &\leq (f,u_\varepsilon^-)-(\nabla u,\nabla u_\varepsilon^-)\\ &=(f+\Delta u,u_\varepsilon^-)\leq \frac{\varepsilon^2}{2}\|f+\Delta u\|^2+\frac{\varepsilon^{-2}}{2}\|u_\varepsilon^-\|^2.\end{aligned}$$

This proves the asserted estimate. □

To define a penalty method that provides a family of approximations $(u_\varepsilon)_{\varepsilon>0}\subset K$, i.e., with $u_\varepsilon\geq\chi$ in Ω for every $\varepsilon>0$, we choose a Lipschitz continuous function $\theta\in W^{1,\infty}(\mathbb{R})$ with

$$\theta'\geq 0,\quad 0\leq\theta\leq 1,\quad \theta(t)=0 \text{ for all } t\leq 0.$$

We assume that $\theta(t)\to 1$ as $t\to+\infty$ with rate $1/t$, i.e., there exists $c_\theta>0$ with

$$1-\theta(t)\leq c_\theta t^{-1}$$

for all $t>0$. Possible choices are $\theta(t)=t/(1+t)$ or $\theta(t)=(2/\pi)\arctan(t)$.

Theorem 5.5 (Conforming penalization) *Let $g=0$ on $\Gamma_{\rm N}$ and define $\theta_\varepsilon(t)=\theta(t/\varepsilon)$ for every $t\in\mathbb{R}$ and $\varepsilon>0$. Let $\xi\in L^2(\Omega)$ be a nonnegative function such that $\xi\geq(-\Delta\chi-f)^+$ and $\xi\geq\partial_n\chi$ on $\Gamma_{\rm N}$ in the sense that*

$$(\nabla\chi,\nabla\phi)-(f,\phi)\leq(\xi,\phi)$$

for all $\phi\in H^1_{\rm D}(\Omega)$ with $\phi\geq 0$ in Ω. There exists a unique function $u_\varepsilon\in H^1_{\rm D}(\Omega)$ such that

$$(\nabla u_\varepsilon,\nabla w)+([\theta_\varepsilon(u_\varepsilon-\chi)-1]\xi,w)=(f,w)$$

for all $w\in H^1_{\rm D}(\Omega)$ that satisfies $u_\varepsilon\geq\chi$ and

$$\|\nabla(u-u_\varepsilon)\|^2\leq\varepsilon c_\theta\|\xi\|_{L^1(\Omega)}$$

Proof Given $\Theta_\varepsilon\in C^1(\mathbb{R})$ with $\Theta_\varepsilon'=\theta_\varepsilon$ we have that Θ_ε is convex and there exists a unique minimizer $u_\varepsilon\in H^1_{\rm D}(\Omega)$ of the functional

$$I_\varepsilon(u)=\frac{1}{2}\int_\Omega|\nabla u|^2\,{\rm d}x+\int_\Omega\Theta_\varepsilon(u-\chi)\,\xi\,{\rm d}x-\int_\Omega(f+\xi)u\,{\rm d}x.$$

The minimizer $u_\varepsilon\in H^1_{\rm D}(\Omega)$ solves the Euler–Lagrange equations

$$(\nabla u_\varepsilon,\nabla w)+([\theta_\varepsilon(u_\varepsilon-\chi)-1]\xi,w)=(f,w)$$

for all $w \in H^1_\mathrm{D}(\Omega)$. Due to the assumption on ξ we have

$$\begin{aligned}\|\nabla(\chi - u_\varepsilon)^+\|^2 &= (\nabla[\chi - u_\varepsilon], \nabla[\chi - u_\varepsilon]^+)\\ &= (\nabla\chi, \nabla[\chi - u_\varepsilon]^+) - (f, (\chi - u_\varepsilon)^+) - (\xi, (\chi - u_\varepsilon)^+)\\ &\quad + (\xi\theta_\varepsilon(u - \chi), (\chi - u_\varepsilon)^+)\\ &\le (\xi\theta_\varepsilon(u_\varepsilon - \chi), (\chi - u_\varepsilon)^+) = 0.\end{aligned}$$

The last identity follows from the fact that almost everywhere in Ω we have either $u_\varepsilon < \chi$; then $\theta_\varepsilon(u_\varepsilon - \chi) = 0$ or $\chi \le u_\varepsilon$ and then $(\chi - u_\varepsilon)^+ = 0$. This proves that $u_\varepsilon \ge \chi$ in Ω and hence $u_\varepsilon \in K$. The variational inequality satisfied by $u \in K$ and the Euler–Lagrange equations fulfilled by $u_\varepsilon \in H^1_\mathrm{D}(\Omega)$ show that

$$\begin{aligned}\|\nabla(u - u_\varepsilon)\|^2 &\le (f, u - u_\varepsilon) + (\xi\theta_\varepsilon(u_\varepsilon - \chi), u - u_\varepsilon) - (f + \xi, u - u_\varepsilon)\\ &= ([\theta_\varepsilon(u_\varepsilon - \chi) - 1]\xi, u - u_\varepsilon) \le ([\theta_\varepsilon(u_\varepsilon - \chi) - 1]\xi, \chi - u_\varepsilon),\end{aligned}$$

where we used $\theta_\varepsilon \le 1$, $\xi \ge 0$, and $u \ge \chi$ in the last estimate. Since

$$s\big(1 - \theta_\varepsilon(s)\big) = \varepsilon(s/\varepsilon)\big(1 - \theta(s/\varepsilon)\big) \le \varepsilon c_\theta$$

we deduce the asserted bound. □

Remark 5.2 The conforming penalization method is practical only if $\chi \in H^2(\Omega)$.

5.1.5 Dual Formulation

We write the obstacle problem as the formally unconstrained minimization of the functional

$$I(u) = \frac{1}{2}\int_\Omega |\nabla u|^2 \,\mathrm{d}x - \int_\Omega f u \,\mathrm{d}x + I_{K_0^+}(u - \chi)$$

with the indicator functional $I_{K_0^+}$ of the set

$$K_0^+ = \{v \in H^1_\mathrm{D}(\Omega) : v \ge 0 \text{ in } \Omega\}.$$

With $\Lambda : H^1_\mathrm{D}(\Omega) \to L^2(\Omega; \mathbb{R}^d)$, $u \mapsto \nabla u$, this can be abstractly written as

$$I(u) = F(\Lambda u) + G(u)$$

and the dual formulation consists in the maximization of (cf. Sect. 4.1.4)

$$p \mapsto D(p) = -F^*(p) - G^*(-\Lambda' p)$$

in the set of functions $p \in L^2(\Omega;\mathbb{R}^d)$. Here, Λ' is the formal adjoint operator $\Lambda' = \nabla' : L^2(\Omega;\mathbb{R}^d) \to H^1_{\mathrm{D}}(\Omega)'$ defined by $\langle \nabla' p, v\rangle = (p, \nabla v)$ for all $v \in H^1_{\mathrm{D}}(\Omega)$. We have $F^* = F$ and for $\mu \in H^1_{\mathrm{D}}(\Omega)'$

$$\begin{aligned} G^*(\mu) &= \sup_{u\in H^1_{\mathrm{D}}(\Omega)} \langle \mu + f, u\rangle - I_{K_0^+}(u-\chi) = \sup_{u\in H^1_{\mathrm{D}}(\Omega)} \langle \mu + f, u+\chi\rangle - I_{K_0^+}(u) \\ &= \langle \mu + f, \chi\rangle + I^*_{K_0^+}(\mu + f) = \langle \mu + f, \chi\rangle + I_{K_0^-}(\mu+f), \end{aligned}$$

where we used a simple calculation to imply $I^*_{K_0^+} = I_{K_0^-}$ with

$$K_0^- = \{\mu \in H^1_{\mathrm{D}}(\Omega)' : \mu \le 0\}.$$

The dual problem thus seeks a maximizing function $p \in L^2(\Omega;\mathbb{R}^d)$ for

$$D(p) = -\frac{1}{2}\int_\Omega |p|^2 \,\mathrm{d}x - \langle \chi, -\nabla' p + f\rangle - I_{K_0^-}(-\nabla' p + f).$$

In the case $\Gamma_{\mathrm{D}} = \partial\Omega$ we have $\nabla' = -\operatorname{div}$. The choice $p = \nabla u$ shows that we have strong duality.

Theorem 5.6 (Strong duality) *Let $\Gamma_{\mathrm{D}} = \partial\Omega$. Then $p = \nabla u$ is maximal for $q \mapsto D(q)$ in the set of functions $q \in L^2(\Omega;\mathbb{R}^d)$ and we have*

$$I(u) = \inf_{v\in H^1_{\mathrm{D}}(\Omega)} I(v) = \sup_{q\in L^2(\Omega;\mathbb{R}^d)} D(q) = D(p).$$

Proof For a functional $\Phi : X \times Y \to \mathbb{R}\cup\{+\infty\}$ and $\overline{x} \in X$ we have

$$\sup_{y\in Y} \Phi(\overline{x}, y) \ge \sup_{y\in Y}\inf_{x\in X} \Phi(x,y),$$

and hence $\inf_{x\in X}\sup_{y\in Y}\Phi(x,y) \ge \sup_{y\in Y}\inf_{x\in X}\Phi(x,y)$. This allows us to deduce that

$$\begin{aligned} I(u) = \inf_{v\in H^1_{\mathrm{D}}(\Omega)} I(v) &= \inf_{v\in H^1_{\mathrm{D}}(\Omega)} \sup_{q\in L^2(\Omega;\mathbb{R}^d)} (q, \nabla v) - \frac{1}{2}\int_\Omega |q|^2\,\mathrm{d}x + G(v) \\ &\ge \sup_{q\in L^2(\Omega;\mathbb{R}^d)} \inf_{v\in H^1_{\mathrm{D}}(\Omega)} -(\operatorname{div} q, v) - \frac{1}{2}\int_\Omega |q|^2\,\mathrm{d}x + G(v) \\ &= \sup_{q\in L^2(\Omega;\mathbb{R}^d)} (-1)\Big(\sup_{v\in H^1_{\mathrm{D}}(\Omega)} (\operatorname{div} q, v) + \frac{1}{2}\int_\Omega |q|^2\,\mathrm{d}x - G(v)\Big) \\ &= \sup_{q\in L^2(\Omega;\mathbb{R}^d)} -G^*(\operatorname{div} q) - \frac{1}{2}\int_\Omega |q|^2\,\mathrm{d}x \\ &= \sup_{q\in L^2(\Omega;\mathbb{R}^d)} D(q). \end{aligned}$$

The direct method in the calculus of variations shows that there exists a unique minimizer for $-D$ in $L^2(\Omega;\mathbb{R}^d)$. For $p=\nabla u$ we have, using $\operatorname{div} p+f\le 0$,

$$\begin{aligned}D(p)&=\frac{1}{2}\int_\Omega |\nabla u|^2\,\mathrm{d}x-(p,\nabla u)-(\chi,\operatorname{div} p+f)\\&=\frac{1}{2}\int_\Omega |\nabla u|^2\,\mathrm{d}x+(\operatorname{div} p+f,u)-(f,u)-(\chi,\operatorname{div} p+f)\\&=\frac{1}{2}\int_\Omega |\nabla u|^2\,\mathrm{d}x-(f,u)+(u-\chi,\operatorname{div} p+f).\end{aligned}$$

The complementarity conditions show that $(u-\chi,\operatorname{div} p+f)=0$ and hence $D(p)=I(u)$. □

5.2 Finite Element Approximation

Given a finite element space that is a subspace of $H^1_\mathrm{D}(\Omega)$, it appears natural to restrict the variational formulation to the discrete space. While this leads to convergence in approximating the exact solution, the practical computation is difficult in general since treating the constraint $u_h\ge\chi$ may be inefficient in practice. The discretization of penalized formulations allows us to use standard solvers for discrete problems but often leads to suboptimal results when the obstacle or the solution is not regular. A more efficient approach is to approximate the obstacle in the finite element space by a function χ_h and to solve the discrete variational inequality by imposing the constraint on the degrees of freedom which often implies that $u_h\ge\chi_h$ holds almost everywhere. We present in this section a priori and a posteriori error estimates for approximating the obstacle problem with finite elements and refer the reader to the textbooks [3, 4] for further details.

5.2.1 *Abstract Error Analysis*

For a Banach space X, a continuous and coercive bilinear form $a: X\times X\to\mathbb{R}$, a closed convex set $K\subset X$, and a bounded linear functional $\ell: X\to\mathbb{R}$, we let $u\in K$ denote the uniquely defined function in K with

$$a(u,v-u)\ge\ell(v-u)$$

for all $v\in K$. For a subspace $X_h\subset X$ and a closed convex set $K_h\subset X_h$, we let $u_h\in K_h$ be the unique solution of

$$a(u_h, v_h - u_h) \geq \ell(v_h - u_h)$$

for all $v_h \in K_h$. Notice that here we do not impose the conformity condition $K_h \subset K$. We define the operator $A : X \to X'$ for every $v \in X$ by

$$\langle Av, w\rangle = a(v, w)$$

for all $w \in X$.

Remark 5.3 If X is a Hilbert space, then the existence and uniqueness of a solution $u \in K$ can be established by showing that the mapping $T_\theta : u \mapsto P_K(u - \theta R(Au - \ell))$ is a contraction for θ sufficiently small. Here, $P_K : X \to K$ is the orthogonal projection onto K and $R : X' \to X$ is the Riesz representation operator.

Theorem 5.7 (Error estimate) *Let H be a Hilbert space such that X is continuously embedded into H and $\langle \phi, v\rangle = (\phi, v)_H$ if $\phi \in H \subset X'$ for all $v \in X$. With the coercivity and continuity constants $\alpha, M > 0$ of a, we have*

$$\frac{\alpha}{2}\|u - u_h\|_X^2 \leq \inf_{v \in K,\, v_h \in K_h} \|Au - \ell\|_{X'}\big(\|u_h - v\|_X + \|u - v_h\|_X\big) + \frac{M^2}{2\alpha}\|u - v_h\|_X^2.$$

If $Au - \ell \in H$, then

$$\frac{\alpha}{2}\|u - u_h\|_X^2 \leq \inf_{v \in K,\, v_h \in K_h} \|Au - \ell\|_H\big(\|u_h - v\|_H + \|u - v_h\|_H\big) + \frac{M^2}{2\alpha}\|u - v_h\|_X^2.$$

Proof The coercivity of a implies that for arbitrary $v \in K$ and $v_h \in K_h$ we have

$$\begin{aligned}
\alpha\|u - u_h\|_X^2 \leq a(u - u_h, u - u_h) &= a(u, u) - a(u, u_h) - a(u_h, u) + a(u_h, u_h) \\
&\leq a(u, v) + \ell(u - v) - a(u, u_h) + a(u_h, v_h) \\
&\quad + \ell(u_h - v_h) - a(u_h, u) \\
&= a(u, v - u_h) + a(u_h, v_h - u) \\
&\quad + \ell(u - v) + \ell(u_h - v_h) \\
&= a(u, v - u_h) + a(u, v_h - u) + \ell(u - v) \\
&\quad + \ell(u_h - v_h) + a(u_h - u, v_h - u) \\
&= \langle Au - \ell, v - u_h\rangle + \langle Au - \ell, v_h - u\rangle \\
&\quad + a(u_h - u, v_h - u).
\end{aligned}$$

This implies the first error estimate. The second estimate follows from the bound

$$\langle Au - \ell, v - u_h\rangle - \langle Au - \ell, v_h - u\rangle \leq \|Au - \ell\|_H\big(\|v - u_h\|_H + \|u - v_h\|_H\big)$$

provided that $Au - \ell \in H$. □

Remark 5.4 If the method is conforming in the sense that $K_h \subset K$, then the terms $\|u_h - v\|_X$ and $\|u_h - v\|_H$ disappear in the estimates.

5.2.2 Application to $P1$-FEM

For a triangulation $\mathscr{T}_h$ of Ω and $\chi \in H^1(\Omega) \cap C(\overline{\Omega})$ with $\chi \le 0$ on $\Gamma_{\rm D}$, we set $\chi_h = \mathscr{I}_h \chi$ and define

$$K_h = \{v_h \in \mathscr{S}^1_{\rm D}(\mathscr{T}_h) : v_h \ge \chi_h\}.$$

The condition $u_h \in K_h$ is, for $u_h \in \mathscr{S}^1_{\rm D}(\mathscr{T}_h)$, equivalent to $u_h(z) \ge \chi_h(z) = \chi(z)$ for all $z \in \mathscr{N}_h$. If χ is not continuous, then a possible definition of the discrete obstacle is $\chi_h = \mathscr{J}_h \chi$ with the Clément interpolant $\mathscr{J}_h : L^1(\Omega) \to \mathscr{S}^1(\mathscr{T}_h)$. Throughout the following we assume that $u \in H^1_{\rm D}(\Omega)$ satisfies $u \ge \chi$ in Ω and

$$(\nabla u, \nabla[v - u]) \ge (f, v - u)$$

for all $v \in H^1_{\rm D}(\Omega)$ with $v \ge \chi$. Correspondingly, we let $u_h \in \mathscr{S}^1_{\rm D}(\mathscr{T}_h)$ be the unique function that satisfies $u_h \ge \chi_h$ and

$$(\nabla u_h, \nabla[v_h - u_h]) \ge (f, v_h - u_h)$$

for all $v_h \in \mathscr{S}^1_{\rm D}(\mathscr{T}_h)$ with $v_h \ge \chi_h$.

Proposition 5.2 (Convergence) *Assume that $\chi_h \to \chi$ in $H^1(\Omega)$ as $h \to 0$. Then we have $u_h \to u$ as $h \to 0$.*

Proof Due to the density of the finite element spaces in $H^1_{\rm D}(\Omega)$, there exists a sequence $(w_h)_{h>0} \subset H^1_{\rm D}(\Omega)$ such that $w_h \in \mathscr{S}^1_{\rm D}(\mathscr{T}_h)$ for every $h > 0$ and $w_h \to u$ in $H^1(\Omega)$ as $h \to 0$. Noting that a standard mollification of the nonnegative function $u - \chi$ provides a nonnegative function $(u - \chi)_\varepsilon = u_\varepsilon - \chi_\varepsilon$ with smooth regularizations u_ε, χ_ε such that $u_\varepsilon \to u$ and $\chi_\varepsilon \to \chi$ in $H^1(\Omega)$ as $\varepsilon \to 0$. We may thus define $v_h = \mathscr{I}_h(u - \chi)_\varepsilon + \chi_h = \mathscr{I}_h u_\varepsilon - \mathscr{I}_h \chi_\varepsilon + \chi_h \in K_h$ as an approximation of u in K_h. An approximation of u_h in K is given by $v = u_h + (\chi - \chi_h)_+ \ge \chi$. For these choices the first estimate of Theorem 5.7 yields the bound

$$\|\nabla(u - u_h)\|^2 \le c\big(\|\nabla(\chi - \chi_h)_+\| + \|\nabla(u - \mathscr{I}_h u_\varepsilon)\| + \|\nabla(\mathscr{I}_h \chi_\varepsilon - \chi_h)\|\big).$$

Inserting χ in the term involving χ_ε shows that the right-hand side converges to 0 as $(\varepsilon, h) \to 0$. □

Theorem 5.8 (Error estimate) *If $\chi, u \in H^2(\Omega)$ and $\chi_h = \mathscr{I}_h \chi$, then*

$$\|\nabla(u - u_h)\|^2 \le ch^2 \|\Delta u + f\| \big(\|D^2 \chi\| + \|D^2 u\|\big) + ch^2 \|D^2 u\|^2.$$

Proof For $v_h = \mathscr{I}_h u$ we have $v_h \ge \chi_h$ and for $v = u_h + (\chi - \chi_h)_+$ we have $v \ge \chi$. Choosing these functions in the second estimate of Theorem 5.7 with $H = L^2(\Omega)$ and incorporating nodal interpolation estimates prove the error estimate. □

5.2.3 A Posteriori Error Analysis

We consider the solutions $u \in K$ and $u_h \in K_h$ of the continuous and discrete obstacle problem with homogeneous boundary conditions on the entire boundary $\Gamma_{\mathrm{D}} = \partial\Omega$, i.e., $u \in K = \{v \in H_0^1(\Omega) : v \ge \chi\}$ satisfies

$$(\nabla u, \nabla[v - u]) \ge (f, v - u)$$

for all $v \in K$, while $u_h \in K_h = \{v_h \in \mathscr{S}_0^1(\mathscr{T}_h) : v_h \ge \chi_h\}$ satisfies for all $v_h \in K_h$

$$(\nabla u_h, \nabla[v_h - u_h]) \ge (f, v_h - u_h).$$

We follow the arguments of [2, 8] and recall the definition of the discrete inner product $(v, w)_h = \int_\Omega \mathscr{I}_h[vw]\,\mathrm{d}x = \sum_{z\in\mathscr{N}_h} \beta_z v(z)w(z)$ for $v, w \in C(\overline{\Omega})$ and $\beta_z = \int_\Omega \varphi_z\,\mathrm{d}x$ for all $z \in \mathscr{N}_h$.

Lemma 5.1 (Discrete Lagrange multiplier) *Let $\lambda_h \in \mathscr{S}_0^1(\mathscr{T}_h)$ be defined by*

$$(\lambda_h, w_h)_h = (f, w_h) - (\nabla u_h, \nabla w_h)$$

for all $w_h \in \mathscr{S}_0^1(\mathscr{T}_h)$. Then we have $\lambda_h \le 0$ and $(\lambda_h, u_h - \chi_h)_h = 0$.

Proof Given $z \in \mathscr{N}_h$ let $\alpha_z \in \mathbb{R}$ be such that $u_h(z) + \alpha_z \ge \chi_h(z)$. The discrete variational inequality with $v_h = u_h + \alpha_z\varphi_z$ and $w_h = \alpha_z\varphi_z$ in the definition of λ_h imply with $\beta_z = \int_\Omega \varphi_z\,\mathrm{d}x > 0$ that

$$\alpha_z\beta_z\lambda_h(z) = \alpha_z\big((f, \varphi_z) - (\nabla u_h, \nabla\varphi_z)\big) \le 0,$$

in particular, $(\nabla u_h, \nabla\varphi_z) - (f, \varphi_z) = 0$ if $u_h(z) > \chi_h(z)$. □

Lemma 5.2 (Attachment) *Let $z \in \mathscr{N}_h \backslash \partial\Omega$ and assume that $w_h \in \mathscr{S}_0^1(\mathscr{T}_h)$ is such that $w_h|_{\omega_z} \le 0$ and $w_h(z) = 0$. Then*

$$\|w_h\|_{L^2(\omega_z)} \le c h_z \sum_{S\in\mathscr{S}_h, z\in S} h_S^{1/2} \|[\![\nabla w_h \cdot n_S]\!]\|_{L^2(S)}.$$

Proof Assume that the right-hand side vanishes. Then $\nabla w_h|_{\omega_z}$ is constant on ω_z and $w_h|_{\omega_z}$ is affine. The conditions $w_h(z) = 0$ and $w_h|_{\omega_z} \le 0$ on ω_z imply that $\|w_h\|_{L^2(\omega_z)} = 0$. Both sides of the asserted estimate define seminorms on $\mathscr{S}_0^1(\mathscr{T}_h)|_{\omega_z}$

and we thus deduce the result if $h_z = 1$. A scaling argument proves the estimate in the general case. □

The lemmas allow us to prove the following error estimate.

Theorem 5.9 (Residual estimate) *Assume that* $\chi_h = \chi$, $\Gamma_{\rm D} = \partial\Omega$, *and let* $f_T = |T|^{-1}\int_T f\,{\rm d}x$ *for every* $T \in \mathcal{T}_h$. *We have*

$$\begin{aligned}(1/c)\|\nabla(u-u_h)\|^2 \le \sum_{T\in\mathcal{T}_h} h_T^2\|f+\Delta u_h-\lambda_h\|_{L^2(T)}^2 + \sum_{T\in\mathcal{T}_h} h_T^2\|f_T-f\|_{L^2(T)}^2 \\ + \sum_{S\in\mathcal{S}_h\cap\Omega} h_S\|[\![\nabla u_h\cdot n_S]\!]\|_{L^2(S)}^2 \\ + \sum_{S\in\mathcal{S}_h^{FB}\cap\Omega} h_S\|[\![\nabla(\chi_h-u_h)\cdot n_S]\!]\|_{L^2(S)}^2\end{aligned}$$

with the skeleton of the discrete free boundary $\mathcal{S}_h^{FB}$ *defined as*

$$\mathcal{S}_h^{FB} = \{S\in\mathcal{S}_h : z_1\in\mathcal{N}_h\cap S,\ u_h(z_1)=\chi_h(z_1),\ z_2\in\mathcal{N}_h\cap\overline{\omega}_{z_1},\ u_h(z_2)>\chi_h(z_2)\}.$$

Proof Abbreviating $e = u - u_h$ and employing the quasi-interpolation operator $\mathcal{J}_h$ we have, since $u_h \in K$,

$$\begin{aligned}\|\nabla e\|^2 &\le (f,e) - (\nabla u_h, \nabla[e-\mathcal{J}_h e]) - (f,\mathcal{J}_h e) + (\lambda_h,\mathcal{J}_h e)_h \\ &= (f, e-\mathcal{J}_h e) - (\nabla u_h, \nabla[e-\mathcal{J}_h e]) + (\lambda_h,\mathcal{J}_h e)_h \\ &= (f-\lambda_h, e-\mathcal{J}_h e) - (\nabla u_h, \nabla[e-\mathcal{J}_h e]) + (\lambda_h, e) \\ &\quad - (\lambda_h,\mathcal{J}_h e) + (\lambda_h,\mathcal{J}_h e)_h.\end{aligned}$$

The first two terms are treated as in the case of the Poisson problem, with an integration by parts, elementwise. The last two terms are controlled with the properties of the discrete inner product by

$$\begin{aligned}(\lambda_h,\mathcal{J}_h e) - (\lambda_h,\mathcal{J}_h e)_h &\le c\sum_{T\in\mathcal{T}_h} h_T^2\|\nabla\lambda_h\|_{L^2(T)}\|\nabla\mathcal{J}_h e\|_{L^2(T)} \\ &\le c\Big(\sum_{T\in\mathcal{T}_h} h_T^4\|\nabla(\lambda_h-f_T)\|_{L^2(T)}^2\Big)^{1/2}\|\nabla\mathcal{J}_h e\| \\ &\le c\Big(\sum_{T\in\mathcal{T}_h} h_T^2\|\lambda_h+\Delta u_h-f_T\|_{L^2(T)}^2\Big)^{1/2}\|\nabla\mathcal{J}_h e\|.\end{aligned}$$

To estimate the term (λ_h, e) let $T \in \mathcal{T}_h$.

(i) If $T \cap \partial\Omega \neq \emptyset$, then with $\omega_T = \cup_{z \in \mathcal{N}_h \cap T} \omega_z$, a Poincaré inequality, and an inverse estimate, we have

$$\int_T e\lambda_h \, \mathrm{d}x \le c h_T^2 \|\nabla \lambda_h\|_{L^2(T)} \|\nabla e\|_{L^2(\omega_T)} \le c h_T \|\lambda_h - f_T\|_{L^2(T)} \|\nabla e\|_{L^2(\omega_T)}.$$

(ii) If $\chi_h < u_h$ or $\chi_h = u_h$ on T, then $\lambda_h = 0$ or $e \ge 0$ on T, respectively, implies

$$\int_T e\lambda_h \, \mathrm{d}x \le 0.$$

(iii) If $T \cap \partial\Omega = \emptyset$ and there exist $z_1, z_2 \in \mathcal{N}_h \cap T$ with $u_h(z_1) = \chi_h(z_1)$ and $u_h(z_2) > \chi_h(z_2)$, then we have $\lambda_h(z_2) = 0$. Using that $0 \le \chi_h - u_h \le u - u_h$, $\lambda_h \le 0$, and $u_h(z_1) - \chi_h(z_1) = 0$, we find with Lemma 5.2 that

$$\begin{aligned} \int_T e\lambda_h \, \mathrm{d}x &\le \int_T (\chi_h - u_h)\lambda_h \, \mathrm{d}x \le h_T \|\nabla \lambda_h\|_{L^2(T)} \|\chi_h - u_h\|_{L^2(T)} \\ &\le c h_T \|\lambda_h - f_T\|_{L^2(T)} \sum_{S \in \mathcal{S}_h, z_1 \in S} h_S^{1/2} \|[\![\nabla(\chi_h - u_h) \cdot n_S]\!]\|_{L^2(S)}. \end{aligned}$$

A combination of the estimates implies the theorem. □

Remark 5.5 A related local lower bound for the error has been derived in [8].

We incorporate an a posteriori error estimate that is based on duality arguments as in the abstract setting of Theorem 4.2. In the stated form it is of limited practical use but reveals that optimality conditions are an important part in a posteriori error estimation. The result does not assume that u_h is a discrete minimizer.

Theorem 5.10 (Functional estimate) *Assume* $\Gamma_{\mathrm{D}} = \partial\Omega$ *and let* $u_h \in K$*. For every* $\widehat{p}_h \in H(\mathrm{div}; \Omega)$ *such that* $\mathrm{div}\, \widehat{p}_h + f \le 0$ *in* Ω*, we have*

$$\|\nabla(u - u_h)\|^2 \le \|\nabla u_h - \widehat{p}_h\|^2 + (\chi - u_h, f + \mathrm{div}\, \widehat{p}_h).$$

Proof The abstract a posteriori error estimate for nonsmooth, strongly convex optimization problems shows, with D as in Theorem 5.6, that

$$\frac{1}{4}\|\nabla(u - u_h)\|^2 \le \frac{1}{2}\big(I(u_h) - D(\widehat{p}_h)\big)$$

for every $\widehat{p}_h \in L^2(\Omega; \mathbb{R}^d)$. With Green's formula, $u_h \in K$, and $\mathrm{div}\, \widehat{p}_h + f \le 0$ we infer

$$
\begin{aligned}
I(u_h) - D(\widehat{p}_h) &= \frac{1}{2}\int_\Omega |\nabla u_h|^2 \,\mathrm{d}x - \int_\Omega f u_h \,\mathrm{d}x + \frac{1}{2}\int_\Omega |\widehat{p}_h|^2 \,\mathrm{d}x \\
&\quad + \int_\Omega \chi(\operatorname{div}\, \widehat{p}_h + f)\,\mathrm{d}x \\
&= \frac{1}{2}\int_\Omega |\nabla u_h|^2 \,\mathrm{d}x - \int_\Omega \widehat{p}_h \cdot \nabla u_h \,\mathrm{d}x + \frac{1}{2}\int_\Omega |\widehat{p}_h|^2 \,\mathrm{d}x \\
&\quad + \int_\Omega (\chi - u_h)(f + \operatorname{div}\, \widehat{p}_h)\,\mathrm{d}x.
\end{aligned}
$$

This proves the error estimate. □

5.3 Iterative Solution Methods

We discuss a locally superlinearly convergent and a globally convergent iteration method. These can be combined to obtain a globally convergent method that has fast convergence properties in an appropriate neighborhood of the discrete solution. The interpretation of the primal-dual active set strategy as a semi-smooth Newton method is due to [5].

5.3.1 Semismooth Newton Iteration

The finite element discretization of the obstacle problem leads to a finite-dimensional minimization problem of the form

$$
\text{Minimize } U \mapsto \frac{1}{2} U^\top A U - B^\top U \quad \text{subject to} \quad U \geq Z
$$

with a positive-definite matrix $A \in \mathbb{R}^{L \times L}$. Here, the vectorial inequality $U \geq Z$ is understood component-wise. Arguing as in the infinite-dimensional situation this is equivalent to finding $(U, \Lambda) \in \mathbb{R}^L \times \mathbb{R}^L$ such that

$$
AU + \Lambda = B, \quad U \geq Z, \quad \Lambda \leq 0, \quad \Lambda_i (U - Z)_i = 0, \; i = 1, 2, \dots, L.
$$

As above we consider the component-wise operation $\min\{0, Y\}$ for a vector $Y \in \mathbb{R}^L$ in the following.

Lemma 5.3 (Complementarity function) *The optimality conditions are satisfied if and only if*

$$AU + \Lambda = B, \quad \mathscr{C}(U, \Lambda) = \Lambda - \min\{0, \Lambda + c(U - Z)\} = 0,$$

where $c > 0$ is an arbitrary positive number.

Proof Suppose that the optimality conditions are satisfied and fix $1 \le i \le L$. If $\Lambda_i = 0$, then $U_i \ge Z_i$ implies $\mathscr{C}_i(U, \Lambda) = 0$. If $\Lambda_i < 0$, then $U_i = Z_i$ and $\Lambda_i - \min\{0, \Lambda_i\} = 0$, i.e., $\mathscr{C}_i(U, \Lambda) = 0$. Assume conversely that $\mathscr{C}(U, \Lambda) = 0$ and fix $1 \le i \le L$. If $\Lambda_i + c(U_i - Z_i) < 0$, then $0 = \mathscr{C}_i(U, \Lambda) = \Lambda_i - \Lambda_i + c(U_i - Z_i) = c(U_i - Z_i)$, i.e., $U_i = Z_i$ and $\Lambda_i < 0$. If $\Lambda_i + c(U_i - Z_i) \ge 0$, then $0 = \mathscr{C}_i(U, \Lambda) = \Lambda_i$ and $U_i \ge Z_i$. □

The lemma motivates defining $F : \mathbb{R}^L \times \mathbb{R}^L \to \mathbb{R}^L \times \mathbb{R}^L$ by

$$F(U, \Lambda) = \begin{bmatrix} F_1(U, \Lambda) \\ F_2(U, \Lambda) \end{bmatrix} = \begin{bmatrix} AU + \Lambda - B \\ \mathscr{C}(U, \Lambda) \end{bmatrix}$$

and compute (U, Λ) with $F(U, \Lambda) = 0$. For a set $\mathscr{A} \subset \{1, 2, \dots, L\}$ we define $I_{\mathscr{A}} \in \mathbb{R}^{L \times L}$ for $1 \le i, j \le L$ by

$$(I_{\mathscr{A}})_{ij} = \begin{cases} 1 & \text{if} \quad i = j \text{ and } i \in \mathscr{A}, \\ 0 & \text{otherwise.} \end{cases}$$

We denote $\mathscr{A}^c = \{1, 2, \dots, L\} \setminus \mathscr{A}$ and note that $I_{\mathscr{A}} + I_{\mathscr{A}^c} = I_L$ is the identity matrix in $\mathbb{R}^{L \times L}$.

Theorem 5.11 (Newton differentiability) *The function F is Newton-differentiable at every $(U, \Lambda) \in \mathbb{R}^L \times \mathbb{R}^L$ and with the set*

$$\mathscr{A} = \{1 \le i \le L : \Lambda_i + c(U_i - Z_i) < 0\}$$

we have

$$DF(U, \Lambda) = \begin{bmatrix} DF_1(U, \Lambda) \\ DF_2(U, \Lambda) \end{bmatrix} = \begin{bmatrix} A & I_L \\ -cI_{\mathscr{A}} & I_{\mathscr{A}^c} \end{bmatrix}.$$

In particular, $DF(U, \Lambda)$ is regular and the semismooth Newton scheme for the iterative solution of $F(U, \Lambda) = 0$ is well-defined and locally superlinearly convergent.

Proof A Newton derivative of the mapping $x \mapsto \min\{x, 0\}$ is given by $x \mapsto \chi_{\mathbb{R}_{<0}}(x)$ with $\chi_{\mathbb{R}_{<0}}(x) = 1$ if $x < 0$, and $\chi_{\mathbb{R}_{<0}}(x) = 0$ if $x \ge 0$; this implies that F is Newton-differentiable with Newton derivative DF. Assume that $DF(U, \Lambda)[V, W]^\top = 0$. Then the equation $DF_2(U, \Lambda)[V, W]^\top = 0$ yields $V|_{\mathscr{A}} = 0$ and $W|_{\mathscr{A}^c} = 0$.

The identity $DF_1(U,\Lambda)[V,W]^\top = 0$ then gives $(AV)|_{\mathscr{A}^c} = 0$, and together with $V|_{\mathscr{A}} = 0$ implies $(AV)^\top V = 0$. Since A is positive definite, we deduce that $V = 0$. From $DF_1(U,\Lambda)[V,W]^\top = 0$ we then also find that $W|_{\mathscr{A}} = 0$. □

One step of the semismooth Newton scheme

$$DF(\widetilde{U},\widetilde{\Lambda})[\delta U,\delta\Lambda]^\top = -F(\widetilde{U},\widetilde{\Lambda})$$

for a given iterate $\widetilde{U},\widetilde{\Lambda} \in \mathbb{R}^L$ is equivalent to

$$\begin{bmatrix} A & I_L \\ -cI_{\widetilde{\mathscr{A}}} & I_{\widetilde{\mathscr{A}}^c} \end{bmatrix} \begin{bmatrix} \delta U \\ \delta\Lambda \end{bmatrix} = - \begin{bmatrix} A\widetilde{U} + \widetilde{\Lambda} - B \\ \widetilde{\Lambda} - \min\{0, \widetilde{\Lambda} + c(\widetilde{U} - Z)\} \end{bmatrix},$$

where $\widetilde{\mathscr{A}} = \{1 \le i \le L : \widetilde{\Lambda}_i + c(\widetilde{U}_i - Z_i) < 0\}$. This system can be written as

$$A(\widetilde{U} + \delta U) + (\widetilde{\Lambda} + \delta\Lambda) = B, \quad (\widetilde{\Lambda} + \delta\Lambda)_{\widetilde{\mathscr{A}}^c} = 0, \quad (\widetilde{U} + \delta U)|_{\widetilde{\mathscr{A}}} = Z|_{\widetilde{\mathscr{A}}}.$$

The semismooth Newton scheme can thus be formulated in the following form which is a version of a primal-dual active set method.

Algorithm 5.1 (*Primal-dual active set method*) Let $(U^0,\Lambda^0) \in \mathbb{R}^L \times \mathbb{R}^L$ and $c > 0$ and compute $(U^k,\Lambda^k)_{k=0,1,\dots}$ via

$$\mathscr{A}_k = \{1 \le i \le L : \Lambda_i^k + c(U_i^k - Z_i) < 0\}$$

and

$$\begin{bmatrix} A & I_L \\ I_{\mathscr{A}_k} & I_{\mathscr{A}_k^c} \end{bmatrix} \begin{bmatrix} U^{k+1} \\ \Lambda^{k+1} \end{bmatrix} = \begin{bmatrix} B \\ I_{\mathscr{A}_k} Z \end{bmatrix}.$$

Stop the iteration if $|||U^{k+1} - U^k||| \le \varepsilon_{\text{stop}}$.

Remarks 5.6 (i) The algorithm converges superlinearly if (U^0,Λ^0) is sufficiently close to the solution (U,Λ), cf. Theorem 4.11. Since the Newton-differentiability only holds in finite-dimensional situations and deteriorates as the dimension increases, the condition on the initial guess becomes more critical for increasing dimensions.
(ii) The degrees of freedom related to the entries $\Lambda|_{\mathscr{A}_k^c}$ can be eliminated from the linear system of equations in the algorithm.
(iii) Since only a finite number of active sets are possible, the algorithm terminates within a finite number of iterations at the exact solution.
(iv) Global convergence of the algorithm and monotonicity $U^{k+1} \ge U^k \ge Z$ for $k \ge 2$ can be proved if A is an M-matrix.
(v) Classical active set strategies define $\mathscr{A}_k = \{1 \le i \le L : U_i^k \le Z_i^k\}$ which corresponds to a formal limit $c \to \infty$.

The MATLAB code displayed in Fig. 5.3 realizes the primal-dual active set method for the obstacle problem. The solution u_h is replaced by the sum $\widetilde{u}_h + \widetilde{u}_{D,h}$ with a

```
function obstacle_newton(d,red)
[c4n,n4e,Db,Nb] = triang_cube(d);
c4n = 3*(c4n-.5); Db = [Db;Nb]; Nb = [];
for j = 1:red
    [c4n,n4e,Db,Nb,P0,P1] = red_refine(c4n,n4e,Db,Nb);
end
nC = size(c4n,1);
u_ini = zeros(nC,1);
dNodes = unique(Db); fNodes = setdiff(1:nC,dNodes)';
[s,m,m_lumped,vol_T] = fe_matrices(c4n,n4e);
[m_Nb,m_lumped_Nb] = fe_matrices_bdy(c4n,Nb);
u_ini(dNodes) = u_D(c4n(dNodes,:));
b = m*f(c4n)+m_Nb*g(c4n)-s*u_ini;
c = 1; I = speye(nC); tz = chi(c4n)-u_ini;
tu_old = zeros(nC,1); tu_new = zeros(nC,1); lambda = zeros(nC,1);
norm_corr = 1; eps_stop = 1E-3;
while norm_corr > eps_stop
    inactive = find(lambda+c*(tu_old-tz)≥0);
    active = setdiff(1:nC,[inactive;dNodes]);
    o = sparse(size(active,2),size(active,2));
    X = [s(fNodes,fNodes),I(active,fNodes)';I(active,fNodes),o];
    x = X\[b(fNodes);tz(active)];
    tu_new(fNodes) = x(1:size(fNodes,1));
    corr = tu_new-tu_old;
    norm_corr = sqrt(corr'*s*corr);
    lambda = zeros(nC,1);
    lambda(active) = x(size(fNodes,1)+(1:size(active,2)));
    tu_old = tu_new;
end
u = tu_new+u_ini;
show_p1(c4n,n4e,Db,Nb,u); drawnow;

function val = chi(x); val = zeros(size(x,1),1);
function val = f(x); val = -2*ones(size(x,1),1);
function val = g(x); val = zeros(size(x,1),1);
function val = u_D(x); r = sqrt(sum(x.^2,2));
val = (r.^2/2-log(r)-1/2);
```

Fig. 5.3 MATLAB implementation of the semismooth Newton method for the obstacle problem

function $\widetilde{u}_{\mathrm{D},h}$ that satisfies Dirichlet boundary conditions. The unknown function $\widetilde{u}_h$ satisfies the constraint $\widetilde{u}_h \geq \widetilde{\chi}_h = \chi_h - \widetilde{u}_{\mathrm{D},h}$ and homogeneous Dirichlet conditions. The function $\widetilde{u}_{\mathrm{D},h}$ also serves as an initial guess for the semismooth Newton iteration.

5.3.2 Global Primal-Dual Method

The discretized obstacle problem can be formulated as a minimization of the mapping

$$u_h \mapsto I(u_h) = F(\nabla u_h) + G(u_h)$$

with the functionals

$$F(\nabla u_h) = \frac{1}{2}\int_\Omega |\nabla u_h|^2 \,\mathrm{d}x$$

and with $f_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ defined through $(f_h, v_h)_h = (f, v_h)$ for all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$,

$$G(u_h) = -(f_h, u_h)_h + I_{K_0^+}(u_h - \chi_h).$$

We equip the space $\mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ and the space of element-wise constant vector fields $\mathscr{L}^0(\mathscr{T}_h)^d$ with the inner products $(\cdot,\cdot)_h$ and $(\cdot,\cdot)$, respectively. This allows us to identify them with their duals. The formal adjoint operator $\nabla' : \mathscr{L}^0(\mathscr{T}_h)^d \to \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ of $\nabla : \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h) \to \mathscr{L}^0(\mathscr{T}_h)^d$ is denoted by $-\operatorname{div}_h^0$ and defined via

$$(-\operatorname{div}_h^0 p_h, v_h)_h = (p_h, \nabla v_h)$$

for all $p_h \in \mathscr{L}^0(\mathscr{T}_h)^d$ and $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$. We define a discrete subdifferential of G by

$$\partial_h G(u_h) = \{v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h) : (v_h, w_h - u_h)_h + G(u_h) \le G(w_h) \text{ f.a. } w_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)\}.$$

Within this setting we reformulate the minimization problem as a saddle-point problem.

Proposition 5.3 (Saddle-point formulation) *The unique minimizer $u_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ of I defines a saddle point $(u_h, \nabla u_h) \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h) \times \mathscr{L}^0(\mathscr{T}_h)^d$ for the functional $L : \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h) \times \mathscr{L}^0(\mathscr{T}_h)^d \to \mathbb{R}$ defined by*

$$L(v_h, q_h) = (q_h, \nabla v_h) - F^*(q_h) + G(v_h),$$

where $F^*(q_h) = (1/2)\int_\Omega |q_h|^2 \,\mathrm{d}x$, i.e., with $p_h = \nabla u_h$ we have

$$L(u_h, q_h) \le L(u_h, p_h) \le L(v_h, p_h)$$

for all $(v_h, q_h) \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h) \times \mathscr{L}^0(\mathscr{T}_h)$.

Proof We first note that for $p_h \in \mathscr{L}^0(\mathscr{T}_h)^d$, we have

$$F(p_h) = \frac{1}{2}\int_\Omega |p_h|^2 \,\mathrm{d}x = \sup_{q_h \in \mathscr{L}^0(\mathscr{T}_h)^d} (p_h, q_h) - F^*(q_h)$$

and $q_h = p_h$ is maximal on the right-hand side. This shows that $L(u_h, p_h) \ge L(u_h, q_h)$ for all $q_h \in \mathscr{L}^0(\mathscr{T}_h)^d$. Since u_h is optimal, i.e., $0 \in \partial F(\nabla u_h) + \partial_h G(u_h)$ and F is strongly convex, we have for all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$

$$\frac{1}{2}\|\nabla(u_h - v_h)\|^2 + I(u_h) \le I(v_h).$$

Therefore, employing $p_h = \nabla u_h$,

$$\begin{aligned} L(v_h, p_h) &= I(v_h) - \frac{1}{2}\|\nabla v_h\|^2 + (\nabla v_h, p_h) - \frac{1}{2}\|p_h\|^2 \\ &= I(v_h) - \frac{1}{2}\|\nabla v_h - p_h\|^2 \ge I(u_h) = L(u_h, p_h). \end{aligned}$$

This proves the proposition. □

The global iterative scheme realizes a subdifferential flow for the functional L; in particular, we use the iteration

$$\begin{aligned} d_t u_h^k &\in -\partial_{v,h} L(u_h^k, \widetilde{p}_h^k) = \operatorname{div}_h^0 \widetilde{p}_h^k - \partial_h G(u_h^k), \\ d_t p_h^k &\in \partial_q L(u_h^k, p_h^k) = \nabla u_h^k - \partial F^*(p_h^k) = \nabla u_h^k - p_h^k, \end{aligned}$$

where the discrete subdifferential $\partial_h G$ is given by

$$\partial_h G(u_h) = -f_h + \partial_h I_{K_0^+}(u_h - \chi_h).$$

The inclusion of the iteration characterizes the unique minimizer $u_h^k \in \mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h)$ of

$$u_h \mapsto \frac{1}{2\tau}\|u_h - u_h^{k-1} - \tau(\operatorname{div}_h^0 \widetilde{p}_h^k + f_h)\|_h^2 + I_{K_0^+}(u_h - \chi_h)$$

in the set of functions $u_h \in \mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h)$. Straightforward considerations imply that

$$u_h^k(z) = \max\{\chi_h(z), u_h^{k-1}(z) + \tau(\operatorname{div}_h^0 \widetilde{p}_h^k + f_h)(z)\}$$

for every $z \in \mathscr{N}_h \setminus \Gamma_{\mathrm{D}}$. The equation for $d_t p_h^k$ of the iteration is equivalent to

$$p_h^k = (1+\tau)^{-1}\big(p_h^{k-1} + \tau \nabla u_h^k\big).$$

The proposed algorithm is a modified version of the abstract primal-dual strategy of Algorithm 4.5 in the sense that the extrapolation is done in the variable p_h in which strong convexity holds.

Algorithm 5.2 (*Primal-dual iteration*) Let $u_h^0 \in \mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h)$, $p_h^0 \in \mathscr{L}^0(\mathscr{T}_h)^d$, and $\tau > 0$ and define $d_t p_h^0 = 0$. Compute the sequences $(u_h^k)_{k=0,1,\dots} \subset \mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h)$ and $(p_h^k)_{k=0,1,\dots} \subset \mathscr{L}^0(\mathscr{T}_h)^d$ via $\widetilde{p}_h^k = p_h^{k-1} + \tau d_t p_h^{k-1}$,

$$u_h^k(z) = \max\{\chi_h(z), u_h^{k-1}(z) + \tau(\operatorname{div}_h^0 \widetilde{p}_h^k + f_h)(z)\}$$

for all $z \in \mathscr{N}_h \setminus \Gamma_{\mathrm{D}}$, and

$$p_h^k = (1+\tau)^{-1}\big(p_h^{k-1} + \tau \nabla u_h^k\big).$$

Stop if $\|d_t p_h^k\| \le \varepsilon_{\text{stop}}$.

The following theorem shows that the iteration converges for every choice of (u_h^0, p_h^0).

Theorem 5.12 (Convergence) *Let $u_h \in \mathscr{S}_{\mathrm{D}}^1(\mathscr{T}_h)$ be the unique minimizer for I_h and set $p_h = \nabla u_h \in \mathscr{L}^0(\mathscr{T}_h)^d$. If $\tau \le ch$ with $c > 0$ sufficiently small, then the iteration of Algorithm 5.2 satisfies*

$$\frac{\tau^2}{2}\sum_{k=1}^{L}\big(\|d_t u_h^k\|_h^2 + \|d_t p_h^k\|^2\big) + \tau \sum_{k=1}^{L}\|p_h - p_h^k\|^2 \le \|p_h - p_h^0\|^2 + \|u_h - u_h^0\|_h^2.$$

In particular, $\|d_t p_h^k\|, \|d_t u_h^k\|_h \to 0$ as $k \to \infty$ and $p_h^k \to p_h$ as $k \to \infty$. Moreover, $u_h^k \to u_h$ as $k \to \infty$.

Proof The inclusion and equation satisfied by the iterates, i.e.,

$$-d_t u_h^k + \operatorname{div}_h^0 \widetilde{p}_h^k \in \partial_h G(u_h^k), \qquad -d_t p_h^k + \nabla u_h^k = p_h^k$$

imply upon testing with $u_h - u_h^k$ and $p_h - p_h^k$, respectively, that

$$\begin{aligned}
&\frac{d_t}{2}\big(\|p_h - p_h^k\|^2 + \|u_h - u_h^k\|_h^2\big) + \frac{\tau}{2}\big(\|d_t u_h^k\|_h^2 + \|d_t p_h^k\|^2\big) + \frac{1}{2}\|p_h - p_h^k\|^2 \\
&\quad = -(d_t p_h^k, p_h - p_h^k) - (d_t u_h^k, u_h - u_h^k)_h + \frac{1}{2}\|p_h - p_h^k\|^2 \\
&\quad \le F^*(p_h) - F^*(p_h^k) - (\nabla u_h^k, p_h - p_h^k) + G(u_h) - G(u_h^k) \\
&\qquad + (-\operatorname{div}_h^0 \widetilde{p}_h^k, u_h - u_h^k)_h \\
&\quad = \big[(\nabla u_h, p_h^k) - F^*(p_h^k) + G(u_h)\big] - \big[(\nabla u_h^k, p_h) - F^*(p_h) + G(u_h^k)\big] \\
&\qquad + (\nabla u_h^k, p_h^k) - (\nabla u_h, p_h^k) + (-\operatorname{div}_h^0 \widetilde{p}_h^k, u_h - u_h^k)_h \\
&\quad = L(u_h, p_h^k) - L(u_h^k, p_h) + (\widetilde{p}_h^k - p_h^k, \nabla[u_h - u_h^k]).
\end{aligned}$$

Since (u_h, p_h) is a saddle-point for L_h, we have $L_h(u_h, p_h^k) \leq L_h(u_h, p_h) \leq L_h(u_h^k, p_h)$. With this and $\widetilde{p}_h^k - p_h^k = -\tau^2 d_t^2 p_h^k$, we deduce that

$$\frac{1}{2}\big(\|p_h - p_h^L\|^2 + \|u_h - u_h^L\|_h^2\big) + \frac{\tau^2}{2}\sum_{k=1}^{L}\big(\|d_t u_h^k\|_h^2 + \|d_t p_h^k\|^2\big) + \frac{\tau}{2}\sum_{k=1}^{L}\|p_h - p_h^k\|^2$$
$$\leq \tau^3 \sum_{k=1}^{L}(-d_t^2 p_h^k, \nabla[u_h - u_h^k]) + \frac{1}{2}\big(\|p_h - p_h^0\|^2 + \|u_h - u_h^0\|_h^2\big).$$

To bound the sum on the right-hand side, we use summation by parts, $d_t p_h^0 = 0$, Young's inequality, and an inverse estimate to verify that

$$\begin{aligned}\tau^3 \sum_{k=1}^{L}(d_t^2 p_h^k, \nabla[u_h - u_h^k]) &= \tau^3 \sum_{k=1}^{L}(d_t p_h^{k-1}, \nabla d_t u_h^k) + \tau^2 (d_t p_h^k, \nabla[u_h - u_h^k])|_{k=0}^{L} \\ &\leq \frac{\tau^2}{4}\Big(\sum_{k=1}^{L} 4\tau^2 \|\nabla d_t u_h^k\|^2 + \|d_t p_h^{k-1}\|^2\Big) \\ &\quad + \frac{\tau^2}{4}\|d_t p_h^L\|^2 + \tau^2\|\nabla(u_h - u_h^L)\|^2 \\ &\leq \frac{\tau^2}{4}\Big(\sum_{k=1}^{L}\|d_t u_h^k\|_h^2 + \|d_t p_h^{k-1}\|^2\Big) + \frac{\tau^2}{4}\|d_t p_h^L\|^2 \\ &\quad + \frac{1}{4}\|u_h - u_h^L\|_h^2,\end{aligned}$$

where we assumed that $\tau \leq ch$ with c such that $2\tau\|\nabla v_h\| \leq \|v_h\|_h$ for all $v_h \in \mathcal{S}_D^1(\mathcal{T}_h)$. This proves the estimate of the theorem and implies that $d_t u_h^k \to 0, d_t p_h^k \to 0$, and $p_h^k \to p_h$ as $k \to \infty$. Since $\nabla u_h^k - p_h^k \to 0$ it follows that $u_h^k \to u_h$. □

The MATLAB code displayed in Fig. 5.4 realizes this scheme. The employed routine `comp_gradient.m` computes the element-wise constant gradient of a $P1$ function. The routine `discrete_divergence.m` provides a matrix that computes $\operatorname{div}_h^0 q_h$ for a vector field $q_h \in \mathcal{L}^0(\mathcal{T}_h)^d$.

```
function obstacle_global(d,red)
[c4n,n4e,Db,Nb] = triang_cube(d);
c4n = 3*(c4n-.5); Db = [Db;Nb]; Nb = [];
for j = 1 : red
    [c4n,n4e,Db,Nb,¬,¬] = red_refine(c4n,n4e,Db,Nb);
end
h = 2^(-red); tau = h/10;
nC = size(c4n,1); nE = size(n4e,1);
dNodes = unique(Db); fNodes = setdiff(1:nC,dNodes)';
[s,m,m_lumped,vol_T] = fe_matrices(c4n,n4e);
D = discrete_divergence(c4n,n4e);
u_old = u_D(c4n); u_new = zeros(nC,1);
p_old = zeros(nE,d); dt_p = zeros(nE,d);
norm_corr = 1; eps_stop = 1E-2;
while norm_corr > eps_stop
    p_tilde = p_old+tau*dt_p;
    div_p_tilde = m_lumped\(D'*reshape(p_tilde',d*nE,1));
    b = u_old+tau*(div_p_tilde+f(c4n));
    u_new(fNodes) = max(chi(fNodes),b(fNodes));
    du_new = comp_gradient(c4n,n4e,u_new);
    p_new = (p_old+tau*du_new)/(1+tau);
    dt_p = (p_new-p_old)/tau; p_old = p_new; u_old = u_new;
    norm_corr = sqrt(vol_T'*sum(dt_p.^2,2))
end
show_p1(c4n,n4e,Db,Nb,u_old); drawnow;

function D = discrete_divergence(c4n,n4e)
[nC,d] = size(c4n); nE = size(n4e,1);
ctr = 0; ctr_max = d*(d+1)*nE;
I = zeros(ctr_max,1); J = zeros(ctr_max,1);
X = zeros(ctr_max,1);
for j = 1 : nE
    X_T = [ones(1,d+1);c4n(n4e(j,:),:)'];
    grads_T = X_T\[zeros(1,d);eye(d)];
    vol_T = det(X_T)/factorial(d);
    for m = 1 : d+1
        for n = 1 : d
            ctr = ctr+1;
            I(ctr) = d*(j-1)+n; J(ctr) = n4e(j,m);
            X(ctr) = -vol_T*grads_T(m,n);
        end
    end
end
D = sparse(I,J,X,d*nE,nC);

function val = f(x); val = -ones(size(x,1),1);
function val = u_D(x); val = zeros(size(x,1),1);
function val = chi(x); val = (-.15)*ones(size(x,1),1);
```

Fig. 5.4 MATLAB implementation of the globally convergent primal-dual method for the obstacle problem

References

1. Brézis, H.R., Stampacchia, G.: Sur la régularité de la solution d'inéquations elliptiques. Bull. Soc. Math. Fr. **96**, 153–180 (1968)
2. Chen, Z., Nochetto, R.H.: Residual type a posteriori error estimates for elliptic obstacle problems. Numer. Math. **84**(4), 527–548 (2000). http://dx.doi.org/10.1007/s002110050009
3. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. Classics in Applied Mathematics, vol. 40. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2002)
4. Glowinski, R., Lions, J.L., Trémolières, R.: Numerical Analysis of Variational Inequalities. Studies in Mathematics and Its Applications, vol. 8. North-Holland, Amsterdam (1981)
5. Hintermüller, M., Ito, K., Kunisch, K.: The primal-dual active set strategy as a semismooth Newton method. SIAM J. Optim. **13**(3), 865–888 (2003). http://dx.doi.org/10.1137/S1052623401383558
6. Kinderlehrer, D., Stampacchia, G.: An Introduction to Variational Inequalities and Their Applications. Classics in Applied Mathematics, vol. 31. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2000)
7. Rodrigues, J.F.: Obstacle Problems in Mathematical Physics. North-Holland Mathematics Studies, vol. 134. North-Holland, Amsterdam
8. Veeser, A.: Efficient and reliable a posteriori error estimators for elliptic obstacle problems. SIAM J. Numer. Anal. **39**(1), 146–167 (2001). http://dx.doi.org/10.1137/S0036142900370812

Chapter 6
The Allen–Cahn Equation

6.1 Analytical Properties

The *Allen–Cahn equation* is a simple mathematical model for certain phase separation processes. It also serves as a prototypical example for semilinear parabolic partial differential equations. The presence of a small parameter that defines the thickness of interfaces separating different phases makes the analysis challenging. Given $u_0 \in L^2(\Omega)$, $\varepsilon > 0$ and $T > 0$, we seek a function $u : [0, T] \times \Omega \to \mathbb{R}$ that solves

$$\partial_t u - \Delta u = -\varepsilon^{-2} f(u), \quad u(0) = u_0, \quad \partial_n u(t, \cdot)|_{\partial\Omega} = 0,$$

for almost every $t \in [0, T]$ and with $f = F'$ for a nonnegative function $F \in C^1(\mathbb{R})$ satisfying $F(\pm 1) = 0$, cf. Fig. 6.1. Unless otherwise stated, we always consider $F(s) = (s^2 - 1)^2/4$ and $f(s) = s^3 - s$ but other choices are possible as well. We always assume that $|u_0(x)| \leq 1$ for almost every $x \in \Omega$. For this model problem we will discuss aspects of its numerical approximation. For further details on modeling aspects and the analytical properties of the Allen–Cahn and other phase-field equations we refer the reader to the textbook [7] and the articles [1, 2, 4, 6, 10, 11].

The Allen–Cahn equation is the L^2-gradient flow of the functional

$$I_\varepsilon(u) = \frac{1}{2} \int_\Omega |\nabla u|^2 \, \mathrm{d}x + \varepsilon^{-2} \int_\Omega F(u) \, \mathrm{d}x.$$

Solutions tend to decrease the energy and develop interfaces separating regions in which it is nearly constant with values close to the minima of F. We refer to the zero level set of the function u as the interface but note that this does not define a sharp separation of the phases. More precisely, the phases are separated by a region of width ε around the zero level set of u often called the *diffuse interface*.

S. Bartels, *Numerical Methods for Nonlinear Partial Differential Equations*,
Springer Series in Computational Mathematics 47,
DOI 10.1007/978-3-319-13797-1_6

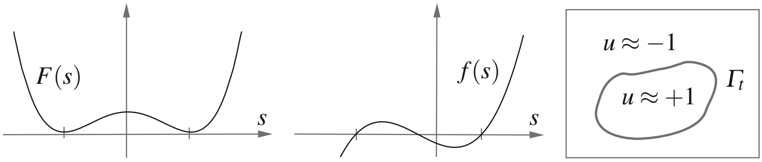


Fig. 6.1 Double well potential $F(s) = (s^2 - 1)^2/4$ and its derivative $f(s) = s^3 - s$ which is monotone outside $[-1, 1]$; solutions develop time-dependent interfaces Γ_t that separate regions in which $u(t, \cdot) \approx \pm 1$

6.1.1 Existence and Regularity

The existence of a unique solution u follows, e.g., from a discretization in time and a subsequent passage to a limit.

Theorem 6.1 (Existence) *For every* $u_0 \in L^2(\Omega)$ *and* $T > 0$ *there exists a weak solution* $u \in H^1([0, T]; H^1(\Omega)') \cap L^2([0, T]; H^1(\Omega))$ *that satisfies* $u(0) = u_0$ *and*

$$\langle \partial_t u, v\rangle + (\nabla u, \nabla v) = -\varepsilon^{-2}(f(u), v)$$

for almost every $t \in [0, T]$ *and every* $v \in H^1(\Omega)$*. If* $u_0 \in H^1(\Omega)$*, then we have* $u \in H^1([0, T]; L^2(\Omega)) \cap L^\infty([0, T]; H^1(\Omega))$ *and*

$$I_\varepsilon(u(T')) + \int_0^{T'} \|\partial_t u\|^2 \,\mathrm{d}t \le I_\varepsilon(u_0)$$

for almost every $T' \in [0, T]$*.*

Proof The existence of a solution follows from an implicit discretization in time that leads to a sequence of well-posed minimization problems. Straightforward a-priori bounds, together with compact embeddings, then show the existence of a weak limit that solves the weak formulation. If $u_0 \in H^1(\Omega)$, then we may formally choose $v = \partial_t u$ to verify that

$$\|\partial_t u\|^2 + \frac{d}{dt}\frac{1}{2}\|\nabla u\|^2 = -\varepsilon^{-2}\frac{d}{dt}\int_\Omega F(u)\,\mathrm{d}x.$$

An integration over $[0, T]$ implies the asserted bound. This procedure can be rigorously carried out for a time-discretized problem, and then the estimate also holds as the time-step size tends to zero. □

Remarks 6.1 (i) Stationary states for the Allen–Cahn equation are the constant functions $u \equiv \pm 1$ and $u \equiv 0$. The state $u \equiv 0$ is unstable.

(ii) For $\Omega = \mathbb{R}^d$ a stationary solution is given by $u(x) = \tanh(x \cdot a/(\sqrt{2}\varepsilon))$ for all $x \in \mathbb{R}^d$ and an arbitrary vector $a \in \mathbb{R}^d$. This characterizes the profile of typical solutions for Allen–Cahn equations across interfaces.

Since the nonlinearity f is monotone outside the interval $[-1, 1]$, solutions of the Allen–Cahn equation satisfy a maximum principle.

Proposition 6.1 (Maximum principle and uniqueness) *If u is a weak solution of the Allen–Cahn equation and $|u_0(x)| \le 1$ for almost every $x \in \Omega$, then $|u(t, x)| \le 1$ for almost every $(t, x) \in [0, T] \times \Omega$. Solutions with this property are unique.*

Proof Let $\widetilde{u} \in H^1([0, T]; H^1(\Omega)') \cap L^2([0, T]; H^1(\Omega))$ be the function obtained by truncating u at ± 1, i.e.,

$$\widetilde{u}(t, x) = \min\{1, \max\{-1, u(t, x)\}\}$$

for almost every $(t, x) \in [0, T] \times \Omega$. Then $\partial_t \widetilde{u} = \partial_t u$, $\nabla \widetilde{u} = \nabla u$, and $f(\widetilde{u}) = f(u)$ in $\{(t, x) \in [0, T] \times \Omega : |\widetilde{u}(t, x)| < 1\}$ and $\partial_t \widetilde{u} = 0$, $\nabla \widetilde{u} = 0$, and $f(\widetilde{u}) = 0$ otherwise. The function $\widetilde{u}$ is therefore a weak solution of the Allen–Cahn equation. If $u - \widetilde{u} \neq 0$, then either $u \ge \widetilde{u} = 1$ and

$$f(u) - f(\widetilde{u}) \ge f'(\widetilde{u})(u - \widetilde{u}) = f'(1)(u - \widetilde{u}) = 2(u - \widetilde{u})$$

or $u \le \widetilde{u} = -1$ and

$$f(u) - f(\widetilde{u}) \le f'(\widetilde{u})(u - \widetilde{u}) = f'(-1)(u - \widetilde{u}) = 2(u - \widetilde{u}).$$

Altogether we find that almost everywhere in $[0, T] \times \Omega$, we have

$$\big(f(u) - f(\widetilde{u})\big)(u - \widetilde{u}) \ge 2|u - \widetilde{u}|^2.$$

The difference $\delta = u - \widetilde{u}$ satisfies

$$(\partial_t \delta, v) + (\nabla \delta, \nabla v) = -\varepsilon^{-2}(f(u) - f(\widetilde{u}), v),$$

and for $v = \delta$, we obtain

$$\frac{1}{2}\frac{d}{dt}\|\delta\|^2 + \|\nabla \delta\|^2 \le -2\varepsilon^{-2}\|\delta\|^2.$$

With $\delta(0) = 0$, it follows directly that $\delta = 0$ in $[0, T] \times \Omega$. If u_1 and u_2 are solutions with $|u_1|, |u_2| \le 1$ in $[0, T] \times \Omega$, then we have

$$|f(u_1) - f(u_2)| \le c_f |u_1 - u_2|$$

almost everywhere in $[0, T] \times \Omega$ with $c_f = \sup_{s \in [-1,1]} |f'(s)|$. The difference $\delta = u_1 - u_2$ satisfies

$$(\partial_t \delta, v) + (\nabla \delta, \nabla v) = -\varepsilon^{-2}(f(u_1) - f(u_2), v)$$

and the choice of $v = \delta$ leads to

$$\frac{1}{2}\frac{d}{dt}\|\delta\|^2 + \|\nabla \delta\|^2 \le c_f \varepsilon^{-2} \|\delta\|^2.$$

An application of Gronwall's lemma implies that $u_1 = u_2$. □

As for the linear heat equation, one can show that the solution is regular. The corresponding bounds depend critically however on the small parameter $\varepsilon > 0$.

Theorem 6.2 (Regularity) *If the Laplace operator is H^2 regular in Ω and $u_0 \in H^1(\Omega)$, then $u \in L^\infty([0,T]; H^2(\Omega)) \cap H^2([0,T]; H^1(\Omega)') \cap H^1([0,T]; H^2(\Omega))$ and there exists $\sigma \ge 0$ such that*

$$\sup_{t\in[0,T]} \|u\|_{H^2(\Omega)} + \Big(\int_0^T \|u_{tt}\|^2_{H^1(\Omega)'} \,\mathrm{d}t\Big)^{1/2} + \Big(\int_0^T \|u_t\|^2_{H^2(\Omega)} \,\mathrm{d}t\Big)^{1/2} \le c\varepsilon^{-\sigma}.$$

If $I_\varepsilon(u_0) \le c$ and $\|\Delta u_0\| \le c\varepsilon^{-2}$, then we may choose $\sigma = 2$.

Proof The proof follows with the arguments that are used to prove the corresponding statements for the linear heat equation, cf. [8]. □

6.1.2 Stability Estimates

In the following stability result we assume that an approximate solution satisfies a maximum principle. This is satisfied for certain numerical approximations and the assumption can be weakened to a uniform L^∞-bound. We recall that Gronwall's lemma states that if a nonnegative function $y \in C([0,T])$ satisfies

$$y(T') \le A + \int_0^{T'} a(t) y(t) \,\mathrm{d}t$$

for all $T' \in [0,T]$ with a nonnegative function $a \in L^1([0,T])$, then we have

$$y(T') \le A \exp\Big(\int_0^T a \,\mathrm{d}t\Big).$$

Together with a Lipschitz estimate, this will be the main ingredient for the following stability result. Due to its exponential dependence on ε^{-2}, it is of limited practical use.

Theorem 6.3 (Stability) *Let* $u \in H^1([0,T]; H^1(\Omega)') \cap L^\infty([0,T]; H^1(\Omega))$ *be a weak solution of the Allen–Cahn equation with* $|u| \le 1$ *almost everywhere in* $[0,T] \times \Omega$. *Let* $\widetilde{u} \in H^1([0,T]; H^1(\Omega)') \cap L^2([0,T]; H^1(\Omega))$ *satisfy* $|\widetilde{u}| \le 1$ *almost everywhere in* $[0,T] \times \Omega$, *and* $\widetilde{u}(0) = \widetilde{u}_0$, *and solve*

$$(\partial_t \widetilde{u}, v) + (\nabla \widetilde{u}, \nabla v) = -\varepsilon^{-2}(f(\widetilde{u}), v) + \langle \widetilde{\mathscr{R}}(t), v \rangle$$

for almost every $t \in [0,T]$, *all* $v \in H^1(\Omega)$, *with a functional* $\widetilde{\mathscr{R}} \in L^2([0,T]; H^1(\Omega)')$. *Then we have*

$$\sup_{t\in[0,T]} \|u - \widetilde{u}\|^2 + \int_0^T \|\nabla(u - \widetilde{u})\|^2 \,\mathrm{d}t$$
$$\le 2\Big(\|u_0 - \widetilde{u}_0\|^2 + \int_0^T \|\widetilde{\mathscr{R}}\|^2_{H^1(\Omega)'} \,\mathrm{d}t\Big) \exp\big((1 + 2c_f\varepsilon^{-2})T\big).$$

Proof With $c_f = \sup_{s\in B_1(0)} |f'(s)|$, we have

$$|f(s_1) - f(s_2)| \le c_f |s_1 - s_2|$$

for all $s_1, s_2 \in \mathbb{R}$. The difference $\delta = u - \widetilde{u}$ satisfies

$$(\partial_t \delta, v) + (\nabla \delta, \nabla v) = -\varepsilon^{-2}\big(f(u) - f(\widetilde{u}), v\big) - \langle \widetilde{\mathscr{R}}, v \rangle$$

for almost every $t \in I$ and every $v \in H^1(\Omega)$. For $v = \delta$ we find that

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\|\delta\|^2 + \|\nabla\delta\|^2 &\le c_f\varepsilon^{-2}\|\delta\|^2 + \|\widetilde{\mathscr{R}}\|_{H^1(\Omega)'}\|\delta\|_{H^1(\Omega)} \\
&\le c_f\varepsilon^{-2}\|\delta\|^2 + \frac{1}{2}\|\widetilde{\mathscr{R}}\|^2_{H^1(\Omega)'} + \frac{1}{2}(\|\delta\|^2 + \|\nabla\delta\|^2) \\
&\le \frac{1}{2}(1 + 2c_f\varepsilon^{-2})\|\delta\|^2 + \frac{1}{2}\|\widetilde{\mathscr{R}}\|^2_{H^1(\Omega)'} + \frac{1}{2}\|\nabla\delta\|^2.
\end{aligned}$$

Absorbing the term $\|\nabla\delta\|^2/2$ on the left-hand side and integrating over $(0, T')$ lead to

$$\|\delta(T')\|^2 + \int_0^{T'} \|\nabla\delta\|^2 \,\mathrm{d}t \le \|\delta(0)\|^2 + \int_0^T \|\widetilde{\mathscr{R}}\|^2_{H^1(\Omega)'} \,\mathrm{d}t + (1 + 2c_f\varepsilon^{-2}) \int_0^{T'} \|\delta\|^2 \,\mathrm{d}t.$$

Defining $A = \|\delta(0)\|^2 + \int_0^T \|\widetilde{\mathscr{R}}\|^2_{H^1(\Omega)'}\,\mathrm{d}t$, $b = (1 + 2c_f\varepsilon^{-2})$, and setting

$$y(t) = \|\delta(t)\|^2 + \int_0^t \|\nabla\delta\|^2\,\mathrm{d}s,$$

we have $y(T') \le A + a\int_0^{T'} y(t)\,\mathrm{d}t$; Gronwall's lemma implies the estimate of the theorem. □

Remark 6.2 The functional $\widetilde{\mathscr{R}}$ models the error introduced by a discretization of the equation so that we may assume that $\|\widetilde{\mathscr{R}}(t)\|^2_{H^1(\Omega)'} \le c\varepsilon^{-\rho}(h^\alpha + \tau^\beta)$ for a mesh-size $h > 0$ and a time-step size $\tau > 0$, and parameters $\alpha, \beta, \rho > 0$. If $\|u_0 - \widetilde{u}_0\|^2 \le h^\gamma$, then we obtain the error estimate

$$\sup_{t\in[0,T]} \|u - \widetilde{u}\|^2 + \int_0^T \|\nabla(u - \widetilde{u})\|^2\,\mathrm{d}t \le c\varepsilon^{-\rho}(h^\alpha + \tau^\beta + h^\gamma)\exp\big((1 + 2c_f\varepsilon^{-2})T\big).$$

Even for the moderate choice $\varepsilon \approx 10^{-1}$, the exponential factor is of the order 10^{40} and it is impossible to compensate this factor with small mesh- and time-step sizes to obtain a useful error estimate. In practice even smaller values of ε are relevant.

To obtain an error estimate that does not depend exponentially on ε^{-1} and which holds without assuming a maximum principle, refined arguments are necessary. The following generalization of Gronwall's lemma allows us to consider a superlinear term.

Proposition 6.2 (Generalized Gronwall lemma) *Suppose that the nonnegative functions $y_1 \in C([0,T])$, $y_2, y_3 \in L^1([0,T])$, $a \in L^\infty([0,T])$, and the real number $A \ge 0$ satisfy*

$$y_1(T') + \int_0^{T'} y_2(t)\,\mathrm{d}t \le A + \int_0^{T'} a(t)y_1(t)\,\mathrm{d}t + \int_0^{T'} y_3(t)\,\mathrm{d}t$$

for all $T' \in [0,T]$. Assume that for $B \ge 0$, $\beta > 0$, and every $T' \in [0,T]$, we have

$$\int_0^{T'} y_3(t)\,\mathrm{d}t \le B\big(\sup_{t\in[0,T']} y_1^\beta(t)\big)\int_0^{T'} (y_1(t) + y_2(t))\,\mathrm{d}t.$$

Set $E = \exp\big(\int_0^T a(t)\,\mathrm{d}t\big)$ and assume that $8AE \le (8B(1+T)E)^{-1/\beta}$. We then have

$$\sup_{t\in[0,T]} y_1(t) + \int_0^T y_2(t)\,\mathrm{d}t \le 8A \exp\Big(\int_0^T a(s)\,\mathrm{d}s\Big).$$

Proof We assume first that $A > 0$, set $\theta = 8AE$, and define

$$I_\theta = \big\{T' \in [0, T] : \Upsilon(T') = \sup_{t\in[0,T']} y_1(t) + \int_0^{T'} y_2(t)\,\mathrm{d}t \le \theta\big\}.$$

Since $y_1(0) \le A < \theta$ and since Υ is continuous and increasing, we have $I_\theta = [0, T_M]$ for some $0 < T_M \le T$. For every $T' \in [0, T_M]$ we have

$$\begin{aligned} y_1(T') + \int_0^{T'} y_2(t)\,\mathrm{d}t &\le A + \int_0^{T'} a(t)y_1(t)\,\mathrm{d}t + B \sup_{t\in[0,T']} y_1^\beta(t) \int_0^{T'} (y_1(t) + y_2(t))\,\mathrm{d}t \\ &\le A + \int_0^{T'} a(t)y_1(t)\,\mathrm{d}t + B(1+T)\theta^{1+\beta}. \end{aligned}$$

An application of the classical Gronwall lemma, the condition on A, and the choice of θ yield that for all $T' \in [0, T_M]$, we have

$$y_1(T') + \int_0^{T'} y_2(t)\,\mathrm{d}t \le (A + B(1+T)\theta^{1+\beta})E \le \frac{\theta}{4}.$$

This implies $\Upsilon(T_M) < \theta$, hence $T_M = T$, and thus proves the lemma if $A > 0$. The argument is illustrated in Fig. 6.2. If $A = 0$, then the above argument holds for every $\theta > 0$ and we deduce that $y_1(t) = y_2(t) = 0$ for all $t \in [0, T]$. □

Remark 6.3 The differential equation underlying the generalized Gronwall lemma has the structure $y' = y^{1+\beta}$. For $\beta > 0$, solutions become unbounded in finite time depending on the initial data, e.g., for $y' = y^2$, we have $y(t) = (t_c - t)^{-1}$ with $t_c = y_0^{-1}$. Therefore, an assumption on A is unavoidable to obtain an estimate on the entire interval $[0, T]$.

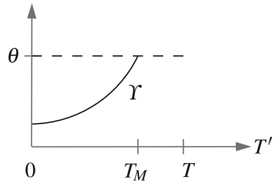


Fig. 6.2 Continuation argument in the proof of the generalized Gronwall lemma

Two elementary properties of the function f are essential for an improved stability result. These define a class of nonlinearities that can be treated with the same arguments.

Lemma 6.1 (Controlled non-monotonicity) *We have* $f'(s) \geq -1$ *and*

$$\big(f(s) - f(r)\big)(s-r) \geq f'(s)(s-r)^2 - 3s(s-r)^3$$

for all $r, s \in \mathbb{R}$.

Proof The lemma follows from the identities $f'(s) = 3s^2 - 1$, $f''(s) = 6s$, and $f'''(s) = 6$ together with a Taylor expansion of f. □

The controlled non-monotonicity of f avoids the use of a Lipschitz estimate. To estimate the resulting term involving f', we employ the smallest eigenvalue of the linearization of the mapping $u \mapsto -\Delta u + f(u(t))$, i.e., of the linear operator $v \mapsto -\Delta v + f'(u(t))v$.

Definition 6.1 For $u \in L^\infty([0,T]; H^1(\Omega))$ let the *principal eigenvalue* $\lambda_{\mathrm{AC}}\colon [0,T] \to \mathbb{R}$ of the linearized Allen–Cahn operator for $t \in [0,T]$ be defined by

$$\lambda_{\mathrm{AC}}(t) = -\inf_{v \in H^1(\Omega)\setminus\{0\}} \frac{\|\nabla v\|^2 + \varepsilon^{-2}\big(f'(u(t))v, v\big)}{\|v\|^2}.$$

Remarks 6.4 (i) As in the theory of ordinary differential equations, the principal eigenvalue contains information about the stability of the evolution.
(ii) If $|u(t)| \leq 1$ in Ω, then we have $-\lambda_{\mathrm{AC}}(t) \geq c_P^2 - 1 - c_f\varepsilon^{-2}$ with the Poincaré constant $c_P = \sup_{v\in H^1(\Omega)\setminus\{0\}} \|v\|/\|v\|_{H^1(\Omega)}$ and $c_f = \sup_{s\in[-1,1]} |f'(s)|$. Therefore, $\lambda_{\mathrm{AC}}(t) \leq 1 + \varepsilon^{-2}$. The evolution is stable as long as $\lambda_{\mathrm{AC}}(t) \leq c$ for an ε-independent constant $c > 0$, and becomes unstable for $\lambda_{\mathrm{AC}}(t) \gg 1$.
(iii) For the stable stationary states $u(t) \equiv \pm 1$, the choice of $v \equiv 1$ shows that we have $\lambda_{\mathrm{AC}}(t) = -2\varepsilon^{-2} \leq 0$, while for the unstable stationary state $u(t) \equiv 0$ we have $\lambda_{\mathrm{AC}}(t) = \varepsilon^{-2}$.
(iv) As long as the curvature of the interface $\Gamma_t = \{x \in \Omega : u(t,x) = 0\}$ is bounded ε-independently, one can show that $\lambda_{\mathrm{AC}}(t)$ is bounded ε-independently, cf. [4].

The generalized Gronwall lemma, the controlled non-monotonicity, and the principal eigenvalue λ_{AC} can be used for an improved stability analysis. We first use the non-monotonicity in the equation for the difference $\delta = u - \widetilde{u}$ tested by δ, i.e.,

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|\delta\|^2 + \|\nabla\delta\|^2 &= -\varepsilon^{-2}\big(f(u) - f(\widetilde{u}), u - \widetilde{u}\big) - \langle\widetilde{\mathscr{R}}, \delta\rangle \\ &\leq -\varepsilon^{-2}\big(f'(u)(u-\widetilde{u}), u - \widetilde{u}\big) \\ &\quad + 3\varepsilon^{-2}\|u\|_{L^\infty(\Omega)}\|u - \widetilde{u}\|^3_{L^3(\Omega)} - \langle\widetilde{\mathscr{R}}, \delta\rangle.\end{aligned}$$

The definition of $\lambda_{\mathrm{AC}}(t)$ implies that

$$-\lambda_{\mathrm{AC}}\|\delta\|^2 \le \|\nabla\delta\|^2 + \varepsilon^{-2}\big(f'(u)\delta, \delta\big)$$

and the combination of the two estimates proves that

$$\frac{1}{2}\frac{d}{dt}\|\delta\|^2 + \|\nabla\delta\|^2 \le \lambda_{\mathrm{AC}}\|\delta\|^2 + \|\nabla\delta\|^2 + 3\varepsilon^{-2}\|u\|_{L^\infty(\Omega)}\|\delta\|^3_{L^3(\Omega)} + \langle\widetilde{\mathscr{R}}, \delta\rangle.$$

By slightly refining the argument we may apply the generalized Gronwall lemma to this equation. In the following theorem we employ the principal eigenvalue defined by an approximate solution to the Allen–Cahn equation. This is in the spirit of a posteriori error estimation to obtain a computable bound for the approximation error. It follows the concept that all information about the problem is extracted from the approximate solution.

Theorem 6.4 (Robust stability) *Let* $0 < \varepsilon \le 1$ *and* $u \in H^1([0,T]; H^1(\Omega)') \cap L^2([0,T]; H^1(\Omega))$ *be the weak solution of the Allen–Cahn equation. Given a function* $\widetilde{u} \in H^1([0,T]; H^1(\Omega)') \cap L^2([0,T]; H^1(\Omega))$ *define* $\widetilde{\mathscr{R}} \in L^2([0,T]; H^1(\Omega)')$ *through*

$$\langle\widetilde{\mathscr{R}}(t), v\rangle = \langle\partial_t\widetilde{u}, v\rangle + (\nabla\widetilde{u}, \nabla v) + \varepsilon^{-2}(f(\widetilde{u}), v)$$

for almost every $t \in [0,T]$ *and all* $v \in H^1(\Omega)$*. Suppose that* $\eta_0, \eta_1 \in L^2([0,T])$ *are such that for almost every* $t \in [0,T]$ *and all* $v \in H^1(\Omega)$*, we have*

$$\langle\widetilde{\mathscr{R}}(t), v\rangle \le \eta_0(t)\|v\| + \eta_1(t)\|\nabla v\|.$$

Assume that $\widetilde{\lambda}_{\mathrm{AC}} \in L^1([0,T])$ *is a function such that for almost every* $t \in (0,T)$*, we have*

$$-\widetilde{\lambda}_{\mathrm{AC}}(t) \le \inf_{v \in H^1(\Omega)\setminus\{0\}} \frac{\|\nabla v\|^2 + \varepsilon^{-2}(f'(\widetilde{u}(t))v, v)}{\|v\|^2},$$

and set $\mu_\lambda(t) = 2\big(2 + (1-\varepsilon^2)\widetilde{\lambda}_{\mathrm{AC}}(t)\big)^+$*. Define*

$$\eta_{\mathrm{AC}}^2 = \|(u-\widetilde{u})(0)\|^2 + \int_0^T (\eta_0^2 + \varepsilon^{-2}\eta_1^2)\,\mathrm{d}t$$

and assume that

$$\eta_{\mathrm{AC}} \le \varepsilon^4\big(6c_S\|\widetilde{u}\|_{L^\infty([0,T];L^\infty(\Omega))}(1+T)\big)^{-1}\Big(8\exp\big(\int_0^T \mu_\lambda(t)\,\mathrm{d}t\big)\Big)^{-3/2},$$

then

$$\sup_{s\in[0,T]} \|u-\widetilde{u}\|^2 + \varepsilon^2 \int_0^T \|\nabla(u-\widetilde{u})\|^2 \,\mathrm{d}t \le 8\eta_{\mathrm{AC}}^2 \exp\Big(\int_0^T \mu_\lambda(t)\,\mathrm{d}t\Big).$$

Proof The difference $\delta = u - \widetilde{u}$ satisfies

$$\langle \partial_t \delta, v\rangle + (\nabla\delta, \nabla v) = -\varepsilon^{-2}(f(u)-f(\widetilde{u}), v) - \langle \widetilde{\mathscr{R}}, v\rangle$$

for almost every $t \in [0,T]$ and all $v \in H^1(\Omega)$. Choosing $v = \delta$, using the assumed bound for $\widetilde{\mathscr{R}}$, noting Lemma 6.1, and using Young's inequality we find

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\|\delta\|^2 + \|\nabla\delta\|^2 &= -\langle\widetilde{\mathscr{R}},\delta\rangle - \varepsilon^{-2}(f(u)-f(\widetilde{u}),\delta)\\
&\le \eta_0\|\delta\| + \eta_1\|\nabla\delta\| - \varepsilon^{-2}(f'(\widetilde{u})\delta,\delta) + 3\varepsilon^{-2}\|\widetilde{u}\|_{L^\infty(\Omega)}\|\delta\|_{L^3(\Omega)}^3\\
&\le \frac{1}{4}\eta_0^2 + \|\delta\|^2 + \frac{\varepsilon^{-2}}{2}\eta_1^2 + \frac{\varepsilon^2}{2}\|\nabla\delta\|^2 - (1-\varepsilon^2)\varepsilon^{-2}(f'(\widetilde{u})\delta,\delta)\\
&\quad - (f'(\widetilde{u})\delta,\delta) + 3\varepsilon^{-2}\|\widetilde{u}\|_{L^\infty(\Omega)}\|\delta\|_{L^3(\Omega)}^3.
\end{aligned}$$

The assumption on $\widetilde{\lambda}_{\mathrm{AC}}(t)$ shows that

$$-\widetilde{\lambda}_{\mathrm{AC}}(t)\|\delta\|^2 \le \|\nabla\delta\|^2 + \varepsilon^{-2}(f'(\widetilde{u})\delta,\delta).$$

Multiplying this estimate by $1-\varepsilon^2$ and using $f'(\widetilde{u}) \ge -1$, we derive the bound

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\|\delta\|^2 + \|\nabla\delta\|^2 &\le \frac{1}{4}\eta_0^2 + \|\delta\|^2 + \frac{\varepsilon^{-2}}{2}\eta_1^2 + \frac{\varepsilon^2}{2}\|\nabla\delta\|^2 + (1-\varepsilon^2)\widetilde{\lambda}_{\mathrm{AC}}\|\delta\|^2\\
&\quad + (1-\varepsilon^2)\|\nabla\delta\|^2 + \|\delta\|^2 + 3\varepsilon^{-2}\|\widetilde{u}\|_{L^\infty(\Omega)}\|\delta\|_{L^3(\Omega)}^3\\
&\le \frac{1}{4}\eta_0^2 + \frac{1}{2}\varepsilon^{-2}\eta_1^2 + (2+(1-\varepsilon^2)\widetilde{\lambda}_{\mathrm{AC}})\|\delta\|^2\\
&\quad + \big(1-\frac{\varepsilon^2}{2}\big)\|\nabla\delta\|^2 + 3\varepsilon^{-2}\|\widetilde{u}\|_{L^\infty(\Omega)}\|\delta\|_{L^3(\Omega)}^3,
\end{aligned}$$

which leads to

$$\frac{d}{dt}\|\delta\|^2 + \varepsilon^2\|\nabla\delta\|^2 \le \eta_0^2 + \varepsilon^{-2}\eta_1^2 + \mu_\lambda\|\delta\|^2 + 6\varepsilon^{-2}\|\widetilde{u}\|_{L^\infty(\Omega)}\|\delta\|_{L^3(\Omega)}^3.$$

Hölder's inequality and the Sobolev estimate $\|v\|_{L^4(\Omega)}^2 \le c_S\|v\|_{H^1(\Omega)}^2$ for $v \in H^1(\Omega)$ yield that

$$\|\delta\|_{L^3(\Omega)}^3 = \int_\Omega |\delta||\delta|^2 \,\mathrm{d}x \le \|\delta\| \|\delta\|_{L^4(\Omega)}^2 \le c_S \|\delta\| (\|\delta\|^2 + \|\nabla \delta\|^2). \tag{6.1}$$

An integration of the last two estimates over $[0, T']$ shows that we are in the situation of Proposition 6.2 with

$$y_1(t) = \|\delta(t)\|^2,\ y_2(t) = \varepsilon^2 \|\nabla \delta(t)\|^2,\ y_3(t) = 6\varepsilon^{-2} \|\widetilde{u}\|_{L^\infty(\Omega)} \|\delta\|_{L^3(\Omega)}^3,$$

and $A = \eta_{\mathrm{AC}}^2$, $B = 6\varepsilon^{-4} \|\widetilde{u}\|_{L^\infty([0,T];L^\infty(\Omega))} c_S$, $\beta = 1/2$, and $E = \exp\big(\int_0^T \mu_\lambda \,\mathrm{d}t\big)$. The proposition thus implies the assertion. □

Remarks 6.5 (i) The robust stability result can be proved for a class of nonlinearities f satisfying the estimates of Lemma 6.1.
(ii) If the exponential factor is bounded by a polynomial in ε^{-1}, then we have improved the stability result of Theorem 6.3. We discuss this question below.

6.1.3 Mean Curvature Flow

The Allen–Cahn equation is closely related to the *mean curvature flow* that seeks for a given hypersurface $\mathscr{M}_0 \subset \mathbb{R}^d$, a family of hypersurfaces $(\mathscr{M}_t)_{t\in[0,T]}$ such that

$$V = -\frac{d-1}{2} H \quad \text{on } \mathscr{M}_t$$

for every $t \in [0, T]$. Here, V is the normal velocity of points on the surface and H is the mean curvature. For a family of spheres $\big((\partial B_{R(t)}(0)\big)_{t\in[0,T]}$ centered at 0 with positive radii $R : [0, T] \to \mathbb{R}$, we have

$$V(t) = R'(t), \quad H(t) = \frac{1}{(d-1)R(t)}.$$

The family of spheres thus solves the mean curvature flow if

$$R' = -\frac{1}{2R},$$

i.e., if $R(t) = (T_c - t)^{1/2}$, where $T_c = R(0)^2$. This equation has a blowup structure and the solution only exists in the interval $[0, T_c)$, cf. Fig. 6.3. To understand the stability of the evolution, we linearize the right-hand side operator $\psi(R) = 1/(2R)$ of the differential equation at the solution $R(t)$ and obtain

$$-\lambda_{\mathrm{MCF}}(t) = \frac{-1}{2R(t)^2} = \frac{-1}{2(T_c - t)}.$$

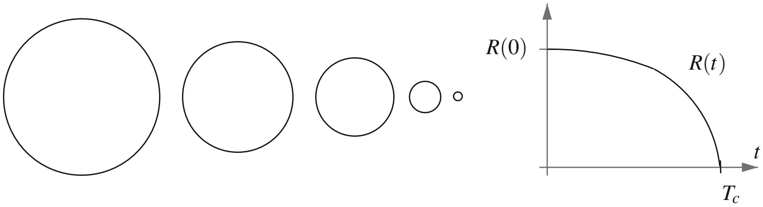


Fig. 6.3 A family of spheres that solve the mean curvature flow within $[0, T_c)$; at $t = T_c$ the surfaces collapse

We thus see that λ_{MCF} is unbounded at $t = T_c$ when the surfaces collapse. This reflects the occurrence of large unbounded normal velocities. Nevertheless, for every $T' < T_c$, we have

$$\int_0^{T'} \lambda_{\mathrm{MCF}}(t)\,\mathrm{d}t = \frac{-1}{2}\big(\log(T_c - T') - \log T_c\big).$$

Assuming that $\lambda_{\mathrm{MCF}} \approx \lambda_{\mathrm{AC}}$, we will deduce below heuristically that the exponential dependence of the stability estimate in Theorem 6.4 is moderate. To understand the relation between the Allen–Cahn equation and the mean curvature flow let $(\mathscr{M}_t)_{t\in[0,T]}$ be a family of surfaces that solve the mean curvature flow. We assume that for every $t \in [0, T]$, we have $\mathscr{M}_t = \partial\Omega_t$ for a simply connected domain $\Omega_t \subset \mathbb{R}^d$ and let $d_{\mathscr{M}}(t, \cdot)$ be the signed distance function to $\mathscr{M}_t$ that is negative inside Ω_t. Given a trajectory $\phi : [0, T] \to \mathbb{R}^d$ of a point $x_0 = \phi(0) \in \mathscr{M}_0$, i.e., we have $\phi(t) \in \mathscr{M}_t$ for all $t \in [0, T]$, its normal velocity given by

$$V(t, \phi(t)) = n(t, \phi(t)) \cdot \phi'(t).$$

Since $d_{\mathscr{M}}(t, \phi(t)) = 0$ for all $t \in [0, T]$ it follows with $n(t, x) = \nabla d_{\mathscr{M}}(t, x)$ for every $x \in \mathscr{M}_t$ that

$$0 = \partial_t d_{\mathscr{M}}(t, \phi(t)) + \nabla d_{\mathscr{M}}(t, \phi(t)) \cdot \phi'(t),$$

i.e., $V(t, x) = -\partial_t d_{\mathscr{M}}(t, x)$ for every $x \in \mathscr{M}_t$. Noting that $D^2 d_{\mathscr{M}} = D(\nabla d_{\mathscr{M}})$ is the shape operator it follows that for the mean curvature we have $(d - 1)H = tr(D^2 d_{\mathscr{M}}) = \Delta d_{\mathscr{M}}$. With $V = -H$ we deduce that $\partial_t d_{\mathscr{M}} - \Delta d_{\mathscr{M}} = 0$ on $\mathscr{M}_t$. The function $\psi(z) = \tanh(z/\sqrt{2})$ satisfies $-\psi''(z) + f(\psi(z)) = 0$, and this implies that for

$$v(t, x) = \psi\Big(\frac{d_{\mathscr{M}}(t, x)}{\varepsilon}\Big)$$

we have

$$\begin{aligned} v_t - \Delta v + \varepsilon^{-2} f(v) &= \varepsilon^{-1}\big(\partial_t d_{\mathscr{M}} - \Delta d_{\mathscr{M}}\big)\,\psi'\big(d_{\mathscr{M}}/\varepsilon\big) - \varepsilon^{-2}\big(\psi''\big(d_{\mathscr{M}}/\varepsilon\big) \\ &\quad + f\big(\psi\big(d_{\mathscr{M}}/\varepsilon\big)\big)\big) \\ &= \varepsilon^{-1}\big(\partial_t d_{\mathscr{M}} - \Delta d_{\mathscr{M}}\big)\,\psi'\big(d_{\mathscr{M}}/\varepsilon\big). \end{aligned}$$

Since $\partial_t d_{\mathscr{M}} - \Delta d_{\mathscr{M}} = 0$ on $\mathscr{M}_t$, we deduce that if $d_{\mathscr{M}}$ is sufficiently smooth, then the function $g = \partial_t d_{\mathscr{M}} - \Delta d_{\mathscr{M}}$ grows linearly in a neighborhood of $\mathscr{M}_t$, i.e., we have $|\partial_t d_{\mathscr{M}} - \Delta d_{\mathscr{M}}| \le c|d_{\mathscr{M}}|$. Noting that the function ψ satisfies $|z\psi'(z)| \le c$, we find that

$$\big|\varepsilon^{-1}(\partial_t d_{\mathscr{M}} - \Delta d_{\mathscr{M}})\psi'\big(d_{\mathscr{M}}/\varepsilon\big)\big| \le c\big|\big(d_{\mathscr{M}}/\varepsilon\big)\psi'\big(d_{\mathscr{M}}/\varepsilon\big)\big| \le c.$$

Therefore, the function $v(t,x) = \psi(d_{\mathscr{M}}(t,x)/\varepsilon)$ solves the dominant terms of the Allen–Cahn equation $\partial_t u - \Delta u = -\varepsilon^{-2} f(u)$ and serves as an approximation of the solution in a neighborhood of width ε of the interface Γ_t. The profile is illustrated in Fig. 6.4. More details can be found in [5].

6.1.4 Topological Changes

The mean curvature flow provides a good approximation of the Allen–Cahn equation in the sense that $v(x,t) = \psi(\mathrm{dist}(x,\mathscr{M}_t)/\varepsilon)$ nearly solves the Allen–Cahn equation; the family $\Gamma_t = \{x \in \Omega : u(x,t) = 0\}$ is a good approximation of a solution for the mean curvature flow. These approximations are valid as long as the interfaces $\mathscr{M}_t$ or Γ_t do not undergo *topological changes*, i.e., as long as $\mathscr{M}_t$ or Γ_t does neither split nor have parts of it disappear. This is closely related to the stability of the solution that is measured by the principal eigenvalue $\lambda_{\mathrm{AC}}(t)$. It can be shown and it follows from the discussion of the mean curvature flow above, that λ_{AC} is bounded from above independently of ε as long as the interface Γ_t is smooth and has bounded curvature. When an interface collapses, large, unbounded velocities occur and the eigenvalue

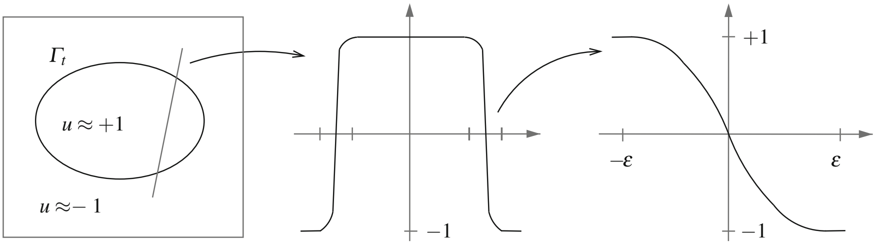


Fig. 6.4 A typical configuration of a solution of the Allen–Cahn equation (*left*) and a solution restricted to a line in the domain (*middle*) together with a magnification of the interface (*right*)

$\lambda_{\rm AC}$ attains the upper bound $\lambda_{\rm AC} \sim \varepsilon^{-2}$. This however only occurs on a time-interval of length comparable to ε^2, the characteristic time scale for the Allen–Cahn equation. Due to this fact, we have for the temporal integral of the principal eigenvalue that occurs in the stability analysis

$$\int_0^T \lambda_{\rm AC}(t)\,\mathrm{d}t \sim 1 + (\text{\# topological changes}) \log(\varepsilon^{-1}).$$

The logarithmic contribution results from the transition regions in which $\lambda_{\rm AC}$ grows like $(T_c - t)^{-1}$ for a topological change at $t = T_c$. Integrating this quantity up to the time $T_c - \varepsilon^2$, where $\lambda_{\rm AC}$ has nearly reached its maximum, reveals that

$$\int_{T_c-1}^{T_c-\varepsilon^2} \lambda_{\rm AC}(t)\,\mathrm{d}t \sim \frac{1}{2} \int_{T_c-1}^{T_c-\varepsilon^2} (T_c - t)^{-1}\,\mathrm{d}t \sim \log(\varepsilon^{-1}).$$

The logarithmic growth in ε^{-1} of the integrated eigenvalue is precisely what is affordable in the estimate of Theorem 6.4 to avoid an exponential dependence on ε^{-1} and instead obtain an algebraic dependence. A typical behavior of the eigenvalue is depicted in Fig. 6.5.

6.1.5 Mass Conservation

The Allen–Cahn equation describes phase transition processes in which the volume fractions of the phases may change and the only stationary configurations represent

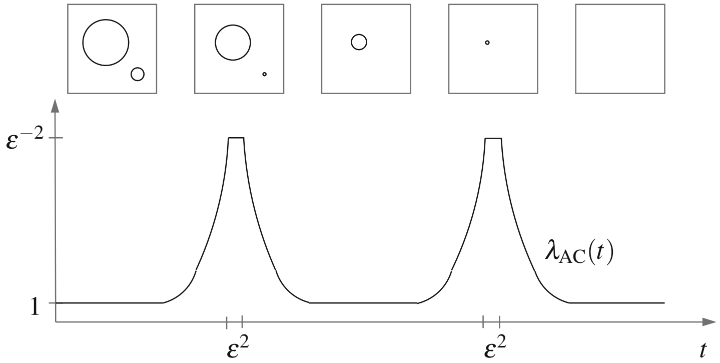


Fig. 6.5 Two topological changes in an evolution defined by the Allen–Cahn equation; the topological changes are accompanied by extreme principal eigenvalues; the eigenvalue increases like $(T_c - t)^{-1}$ before a topological change occurs at T_c

single phases. This corresponds, e.g., to melting processes. In order to model phase separation processes in which the volume fractions are preserved, a constraint has to be incorporated or a fourth order evolution has to be considered. The latter is the H^{-1}-gradient flow of the energy I_ε, where $H^{-1}(\Omega) = X_0'$ is the dual of the space $X_0 = \{v \in H^1(\Omega) : \int_\Omega v \,\mathrm{d}x = 0\}$, i.e.,

$$(\partial_t u, v)_{-1} = -(\nabla u, \nabla v) - \varepsilon^{-2}(f(u), v).$$

Here, the inner product $(v, w)_{-1}$ is for $v, w \in H^{-1}(\Omega)$ defined by

$$(v, w)_{-1} = \int_\Omega \nabla(-\Delta^{-1} v) \cdot \nabla(-\Delta^{-1} w) \,\mathrm{d}x,$$

where $-\Delta^{-1} v$ and $\Delta^{-1} w \in X_0$ are the unique solutions of the Poisson problem

$$-\Delta u = f \text{ in } \Omega, \quad \partial_n u|_{\partial\Omega} = 0$$

with vanishing mean for the right-hand sides $f = v$ and $f = w$, respectively. In the strong form the gradient flow reads

$$\partial_t u = -\Delta\phi, \ \phi = \Delta u - \varepsilon^{-2} f(u),$$

together with homogeneous Neumann boundary conditions on $\partial\Omega$ for u and ϕ and initial conditions for u. The variable ϕ is the *chemical potential* and the system is called the *Cahn–Hilliard equation* which can be analyzed with the techniques discussed above. Mass conservation is a consequence of the fact that $\partial_t u$ has vanishing integral mean. Solutions do not obey a maximum principle but satisfy certain L^∞-bounds.

6.2 Error Analysis

In this section we discuss error estimates for numerical approximations of the Allen–Cahn equation obtained with the implicit Euler scheme. The stability result of Theorem 6.4 is already formulated in the spirit of an a posteriori error analysis. We discuss results from [3, 8, 9].

6.2.1 *Residual Estimate*

We include an estimate for the residual of an approximation obtained with the implicit Euler scheme. The result can be modified to control the error of other approximation schemes.

Proposition 6.3 (Residual bounds) *Let* $0 = t_0 < t_1 < \cdots < t_K \leq T$ *and* $\tau_k = t_k - t_{k-1}$, $k = 1, 2, \ldots, K$, *and* $(\mathscr{T}_k)_{k=0,\ldots,K}$ *a sequence of regular triangulations of* Ω. *Suppose that* $(u_h^k)_{k=0,\ldots,K} \subset H^1(\Omega)$, *for* $k = 1, 2, \ldots, K$ *and all* $v_h \in \mathscr{S}^1(\mathscr{T}_k)$, *satisfies*

$$\tau_k^{-1}(u_h^k - \mathscr{I}_k u_h^{k-1}, v_h) + (\nabla u_h^k, \nabla v_h) = -\varepsilon^{-2}(f(u_h^k), v_h),$$

where $\mathscr{I}_k$ *denotes the nodal interpolation operator related to* $\mathscr{S}^1(\mathscr{T}_k)$. *Let* $u_{h,\tau} \in H^1([0,T]; H^1(\Omega))$ *be the piecewise linear interpolation in time of* $(u_h^k)_{k=0,\ldots,K}$ *and define* $\mathscr{R} \in L^2(I; H^1(\Omega)')$ *for* $t \in [0,T]$ *and* $v \in H^1(\Omega)$ *by*

$$\langle \mathscr{R}(t), v\rangle = (\partial_t u_{h,\tau}, v) + (\nabla u_{h,\tau}, \nabla v) + \varepsilon^{-2}(f(u_{h,\tau}), v).$$

For almost every $t \in [t_{k-1}, t_k]$ *and all* $v \in H^1(\Omega)$ *we have*

$$\langle \mathscr{R}(t), v\rangle \leq (\eta^k_{\text{time}'} + \eta^k_{\text{coarse}})\|v\| + (C_{C\ell}\eta^k_{\text{space}} + \eta^k_{\text{time}})\|\nabla v\|,$$

where $\rho_k = \|u_h^k\|_{L^\infty(\Omega)} + \|u_h^{k-1}\|_{L^\infty(\Omega)}$,

$$\begin{aligned}\eta^k_{\text{space}} = &\Big(\sum_{T \in \mathscr{T}_h^k} h_T^2 \|\tau_k^{-1}(u_h^k - \mathscr{I}_k u_h^{k-1}) - \Delta_{\mathscr{T}_k} u_h^k + \varepsilon^{-2} f(u_h^k)\|^2_{L^2(T)}\Big)^{1/2} \\ &+ \Big(\sum_{S \in \mathscr{S}_h^k \cap \Omega} h_S \|[\![\nabla u_h^k \cdot n_S]\!]\|^2_{L^2(S)}\Big)^{1/2} + \Big(\sum_{S \in \mathscr{S}_h^k \cap \partial\Omega} h_S \|\nabla u_h^k \cdot n\|^2_{L^2(S)}\Big)^{1/2},\end{aligned}$$

and

$$\begin{aligned}\eta^k_{\text{time}'} &= \varepsilon^{-2}\|f'\|_{L^\infty(B_{\rho_k})}\|u_h^{k-1} - u_h^k\|, \\ \eta^k_{\text{time}} &= \|\nabla(u_h^{k-1} - u_h^k)\|, \\ \eta^k_{\text{coarse}} &= \tau_k^{-1}\|\mathscr{I}_k u_h^{k-1} - u_h^{k-1}\|.\end{aligned}$$

Proof For almost every $t \in (t_{k-1}, t_k)$, $k = 1, 2, \ldots, K$, and all $v \in H^1(\Omega)$, we have by definition of $\mathscr{R}$ that

$$\begin{aligned}\langle \mathscr{R}(t), v\rangle &= \tau_k^{-1}(u_h^k - u_h^{k-1}, v) + (\nabla u_{h,\tau}(t), \nabla v) + \varepsilon^{-2}(f(u_{h,\tau}(t)), v) \\ &= \tau_k^{-1}(u_h^k - \mathscr{I}_k u_h^{k-1}, v) + (\nabla u_h^k, \nabla v) + \varepsilon^{-2}(f(u_h^k), v) \\ &\quad + (\nabla(u_{h,\tau}(t) - u_h^k), \nabla v) + \varepsilon^{-2}(f(u_{h,\tau}(t)) - f(u_h^k), v) \\ &\quad + \tau_k^{-1}(\mathscr{I}_k u_h^{k-1} - u_h^{k-1}, v) \\ &= I + II + \cdots + VI.\end{aligned}$$

Since the sum of the first three terms vanishes for all $v \in \mathscr{S}^1(\mathscr{T}_k)$, we may insert the Clément interpolant $\mathscr{J}_k v \in \mathscr{S}^1(\mathscr{T}_k)$ of v. An element-wise integration by parts and estimates for the Clément interpolant lead to

$$I + II + III = \langle r_h^k, v - \mathscr{J}_k v \rangle \le C_{C\ell} \eta_{\text{space}}^k \|\nabla v\|.$$

A repeated application of Hölder's inequality, the identity

$$f(u_{h,\tau}(t)) - f(u_h^k) = \Big(\int_0^1 f'(r u_{h,\tau}(t) + (1-r) u_h^k) \,\mathrm{d}r \Big)(u_{h,\tau}(t) - u_h^k),$$

and the linearity of $u_{h,\tau}$ in t lead to

$$\begin{aligned} IV + V &\le \|\nabla(u_{h,\tau}(t) - u_h^k)\| \|\nabla v\| + \varepsilon^{-2} \|f'\|_{L^\infty(B_{\rho_k})} \|u_{h,\tau}(t) - u_h^k\| \|v\| \\ &\le \eta_{\text{time}'}^k \|v\| + \eta_{\text{time}}^k \|\nabla v\|. \end{aligned}$$

A further application of Hölder's inequality proves

$$VI \le \tau_k^{-1} \|\mathscr{I}_k u_h^{k-1} - u_h^{k-1}\| \|v\| = \eta_{\text{coarse}}^k \|v\|.$$

A combination of the estimates leads to the asserted bound. □

In combination with Theorem 6.4 we obtain the following a posteriori error estimate. It bounds the approximation error in terms of computable quantities provided that the error estimator is sufficiently small and depends exponentially only on the temporal average of the principal eigenvalue defined by the numerical approximation.

Theorem 6.5 (A posteriori error estimate) *Assume that we are in the setting of Proposition 6.3 and suppose that $\lambda_{\text{AC}}^h \in L^1([0,T])$ is a function, such that for almost every $t \in (0,T)$, we have*

$$-\lambda_{\text{AC}}^h(t) \le \inf_{v \in H^1(\Omega)\setminus\{0\}} \frac{\|\nabla v\|^2 + \varepsilon^{-2}(f'(u_{h,\tau}(t))v, v)}{\|v\|^2},$$

and set $\mu_\lambda(t) = 2(2 + (1-\varepsilon^2)\lambda_{\text{AC}}^h(t))^+$. Define $\eta_\ell(t) = \eta_\ell^k$ for $t \in (t_{k-1}, t_k)$, $k = 1, 2, \dots, K$, and $\ell \in \{\text{time}', \text{time}, \text{space}, \text{coarse}\}$ and let

$$\eta_{\text{AC}}^2 = \|(u - u_h^0)(0)\|^2 + \int_0^T (\eta_{\text{time}'}^2 + \eta_{\text{coarse}}^2 + \varepsilon^{-2}\eta_{\text{time}}^2 + \varepsilon^{-2}\eta_{\text{space}}^2) \,\mathrm{d}t.$$

If

$$\eta_{\text{AC}} \le \varepsilon^4 \big(6 c_S \|u_{h,\tau}\|_{L^\infty([0,T];L^\infty(\Omega))}(1+T)\big)^{-1} \Big(8 \exp\Big(\int_0^T \mu_\lambda(t)\,\mathrm{d}t\Big)\Big)^{-3/2},$$

then we have

$$\sup_{s\in[0,T]} \|u-u_{h,\tau}\|^2 + \varepsilon^2 \int_0^T \|\nabla(u-u_{h,\tau})\|^2 \,\mathrm{d}t \le 8\eta^2 \exp\Big(\int_0^T \mu_\lambda(t)\,\mathrm{d}t\Big).$$

Proof The theorem is an immediate consequence of Proposition 6.3 and Theorem 6.4. □

6.2.2 A Priori Error Analysis

To derive a robust a priori error estimate for a semidiscrete in time approximation scheme, we try to follow the arguments used in the stability analysis of Theorem 6.3 with exchanged roles of the exact solution and its numerical approximation. As above we avoid the use of a Lipschitz estimate for the nonlinearity, and instead employ a linearization. The non-monotonicity of the resulting equation is controlled by a cubic term. The linearization allows us to incorporate the principal eigenvalue that is assumed to be well-behaved in the sense that a discrete integral grows only logarithmically in ε^{-1}.

Proposition 6.4 (Discrete stability) *Given $\tau > 0$ let $(U^k)_{k=0,\dots,K} \subset H^1(\Omega)$ be such that*

$$(d_t U^k, v) + (\nabla U^k, \nabla v) = -\varepsilon^{-2}(f(U^k), v)$$

for $k = 1, 2, \dots, K$ and all $v \in H^1(\Omega)$. We then have

$$I_\varepsilon(u^L) + (2 - 2\tau\varepsilon^{-2})\frac{\tau}{2}\sum_{k=1}^K \|d_t U^k\|^2 \le I_\varepsilon(u^0)$$

for every $1 \le L \le K$. Moreover, if $\|U^0\|_{L^\infty(\Omega)} \le 1$, then $\|U^k\|_{L^\infty(\Omega)} \le 1$ for $k = 1, 2, \dots, K$.

Proof The mean value theorem shows that for every $x \in \Omega$ there exists a number ξ_x such that

$$f(U^k)d_t U^k = d_t F(U^k) + \frac{\tau}{2} f'(\xi_x)(d_t U^k)^2.$$

Using that $f'(\xi_x) \ge -1$ and choosing $v = d_t U^k$, we deduce that

$$\|d_t U^k\|^2 + \frac{d_t}{2}\|\nabla U^k\|^2 + \frac{\tau}{2}\|\nabla d_t U^k\|^2 + d_t \varepsilon^{-2} \int_\Omega F(U^k)\,\mathrm{d}x - \frac{\tau\varepsilon^{-2}}{2}\|d_t U^k\|^2 \le 0.$$

Multiplication by τ and summation over $k = 1, 2, \ldots, L$ imply the assertion. A truncation argument and the characterization of U^k as the minimum of a functional I_ε^k show that $\|U^k\|_{L^\infty(\Omega)} \leq 1$ provided that U^0 has this property. □

Proposition 6.5 (Consistency) *Assume that the weak solution of the Allen–Cahn equation satisfies $u \in C([0, T]; H^1(\Omega))$ and $u \in H^2([0, T]; H^1(\Omega)')$ with*

$$\int_0^T \|u_{tt}\|^2_{H^1(\Omega)'} \,\mathrm{d}t \leq c\varepsilon^{-2\sigma}.$$

For $u^k = u(t_k)$, $k = 0, 1, \ldots, K$, we have

$$(d_t u^k, v) + (\nabla u^k, \nabla v) = -\varepsilon^{-2}(f(u^k), v) + \mathscr{C}_\tau(t_k; v)$$

for all $v \in H^1(\Omega)$ with consistency functionals $\mathscr{C}_\tau(t_k)$ satisfying

$$\tau \sum_{k=1}^K \|\mathscr{C}_\tau(t_k)\|^2_{H^1(\Omega)'} \leq c\tau^2\varepsilon^{-2\sigma}.$$

We have $\sigma = 2$ if $I(u_0) \leq c$.

Proof Noting that

$$(d_t u^k, v) + (\nabla u^k, \nabla v) + \varepsilon^{-2}(f(u^k), v) = (d_t u^k - \partial_t u(t_k), v) = \mathscr{C}_\tau(t_k; v)$$

for all $v \in H^1(\Omega)$, arguing as in the case of the linear heat equation, and incorporating Theorem 6.2 proves the asserted bound. □

The following lemma is a generalization of the classical discrete Gronwall lemma which states that if

$$y^{L'} \leq A + \tau \sum_{k=1}^{L'} a_k y^k$$

for $0 \leq L' \leq L$ and if $\tau a_k \leq 1/2$ for $k = 1, 2, \ldots, L$, then we have

$$\sup_{k=0,\ldots,L} y^k \leq 2A \exp\Big(2\tau \sum_{k=1}^L a_k\Big).$$

The condition $a_k\tau \leq 1/2$ is required to absorb the term $a_{L'}y^{L'}$.

Lemma 6.2 (Generalized discrete Gronwall lemma) *Let $\tau > 0$ and suppose that the nonnegative real sequences $(y_\ell^k)_{k=0,\ldots,K}$, $\ell = 1, 2, 3$, $(a_k)_{k=0,\ldots,K}$, and the real number $A \geq 0$ satisfy*

$$y_1^L + \tau \sum_{k=1}^{L} y_2^k \le A + \tau \sum_{k=1}^{L} a_k y_1^k + \tau \sum_{k=1}^{L-1} y_3^k$$

for all $L = 0, 1, \dots, K$, $\sup_{k=1,\dots,K} \tau a_k \le 1/2$, *and* $K\tau \le T$. *Assume that for* $B \ge 0$, $\beta > 0$, *and every* $L = 1, 2, \dots, K$, *we have*

$$\tau \sum_{k=1}^{L-1} y_3^k \le B\big(\sup_{k=1,\dots,L-1} (y_1^k)^\beta \big) \tau \sum_{k=1}^{L-1} (y_1^k + y_2^k).$$

Set $E = \exp\big(2\tau \sum_{k=1}^{K} a_k\big)$ *and assume that* $8AE \le (8B(1+T)E)^{-1/\beta}$. *Then*

$$\sup_{k=0,\dots,K} y_1^k + \tau \sum_{k=1}^{K} y_2^k \le 8A \exp\Big(2\tau \sum_{k=1}^{K} a_k\Big).$$

Proof Set $\theta = 8AE$. We proceed by induction and suppose that

$$\sup_{k=0,\dots,L-1} y_1^k + \tau \sum_{k=1}^{L-1} y_2^k \le \theta.$$

This is satisfied for $L = 1$. For every $L' = 1, 2, \dots, L$, we then have due to the assumptions of the lemma that

$$\begin{aligned} y_1^{L'} + \tau \sum_{k=1}^{L'} y_2^k &\le A + \tau \sum_{k=1}^{L'} a_k y_1^k + B\big(\sup_{k=1,2,\dots,L'-1} (y_1^k)^\beta \big) \tau \sum_{k=1}^{L'-1} (y_1^k + y_2^k) \\ &\le A + \tau \sum_{k=1}^{L'} a_k y_1^k + B(1+T)\theta^{1+\beta}. \end{aligned}$$

The classical discrete Gronwall lemma, the condition on A, and the estimate $\theta^\beta \le (8B(1+T)E)^{-1}$ prove that for all $L' = 1, 2, \dots, L$, we have

$$y_1^{L'} + \tau\varepsilon^2 \sum_{k=1}^{L'} y_2^k \le 2(A + B(1+T)\theta^{1+\beta})E \le \frac{\theta}{2}.$$

This completes the inductive argument and proves the lemma. □

The a priori bounds and the generalized discrete Gronwall lemma lead to a robust a priori error estimate under an assumption on the principal eigenvalue that is motivated by analytical considerations.

Theorem 6.6 (A priori error estimate) *Assume $\varepsilon \le 1$, $I_\varepsilon(u_0) \le c_0$, and that there are $c_1 > 0$, $\kappa \ge 0$ with*

$$\tau \sum_{k=1}^{K} \lambda_{\mathrm{AC}}^{+}(t_k) \le c_1 + \log \varepsilon^{-\kappa}.$$

Then there exists a constant $c_2 > 0$ such that if $\tau \le c_2 \varepsilon^{7+6\kappa}$, we have

$$\sup_{k=1,\dots,K} \|u(t_k) - U^k\|^2 + \tau\varepsilon^2 \sum_{k=1}^{K} \|\nabla(u(t_k) - U^k)\|^2 \le c\tau^2\varepsilon^{-6-4\kappa}.$$

Proof Denoting $u^k = u(t_k)$ the error $e^k = u^k - U^k$ satisfies the identity

$$(d_t e^k, v) + (\nabla e^k, \nabla v) = -\varepsilon^{-2}(f(u^k) - f(U^k), v) + \mathscr{C}_\tau(t_k, v)$$

for all $v \in H^1(\Omega)$. Lemma 6.1, the definition of $\lambda_{\mathrm{AC}}(t_k)$, and $\|u^k\|_{L^\infty(\Omega)} \le 1$ imply that

$$\begin{aligned}
-\varepsilon^{-2}(f(u^k) - f(U^k), e^k) &\le -\varepsilon^{-2}(f'(u^k)e^k, e^k) + 3\varepsilon^{-2}\|u^k\|_{L^\infty(\Omega)}\|e^k\|^3_{L^3(\Omega)} \\
&= -(1-\varepsilon^2)\varepsilon^{-2}(f'(u^k)e^k, e^k) - (f'(u^k)e^k, e^k) \\
&\quad + 3\varepsilon^{-2}\|e^k\|^3_{L^3(\Omega)} \\
&\le (1-\varepsilon^2)\lambda_{\mathrm{AC}}(t_k)\|e^k\|^2 + (1-\varepsilon)\|\nabla e^k\|^2 \\
&\quad + \|e^k\|^2 + 3\varepsilon^{-2}\|e^k\|^3_{L^3(\Omega)}.
\end{aligned}$$

Hence, for the choice of $v = e^k$, we find that

$$\begin{aligned}
\frac{1}{2} d_t\|e^k\|^2 + \frac{\tau}{2}\|d_t e^k\|^2 + \|\nabla e^k\|^2 &= \mathscr{C}_\tau(t_k, e^k) - \varepsilon^{-2}(f(u^k) - f(U^k), e^k) \\
&\le \frac{\varepsilon^{-2}}{2}\|\mathscr{C}_\tau(t_k)\|^2_{H^1(\Omega)'} + \frac{\varepsilon^2}{2}\|e^k\|^2 + \frac{\varepsilon^2}{2}\|\nabla e^k\|^2 \\
&\quad + (1-\varepsilon^2)\lambda_{\mathrm{AC}}(t_k)\|e^k\|^2 + (1-\varepsilon^2)\|\nabla e^k\|^2 \\
&\quad + \|e^k\|^2 + 3\varepsilon^{-2}\|e^k\|^3_{L^3(\Omega)}.
\end{aligned}$$

Using $(a+b)^3 \le 4(a^3 + b^3)$ and $\tau\|d_t e^k\|_{L^\infty(\Omega)} \le 4$ we find that

$$\|e^k\|^3_{L^3(\Omega)} \le 4\big(\|e^{k-1}\|^3_{L^3(\Omega)} + \tau^3\|d_t e^k\|^3_{L^3(\Omega)}\big) \le 4\|e^{k-1}\|^3_{L^3(\Omega)} + 16\tau^2\|d_t e^k\|^2.$$

If τ is sufficiently small so that $48\tau\varepsilon^{-2} \le 1/2$, then the combination of the last two estimates implies

$$d_t\|e^k\|^2 + \varepsilon^2\|\nabla e^k\|^2 \le \varepsilon^{-2}\|\mathscr{C}_\tau(t_k)\|^2_{H^1(\Omega)'} + \mu_\lambda^k\|e^k\|^2 + 48\varepsilon^{-2}\|e^{k-1}\|^3_{L^3(\Omega)}, \quad (6.2)$$

where $\mu_\lambda^k = 2(2 + \lambda_{\mathrm{AC}}^+(t_k))$. We set

$$y_1^k = \|e^k\|^2, \; y_2^k = \varepsilon^2 \|\nabla e^k\|^2, \; y_3^k = 48\varepsilon^{-2}\|e^k\|^3_{L^3(\Omega)}.$$

Noting that $e^0 = 0$ and

$$\|e^{k-1}\|^3_{L^3(\Omega)} \le \|e^{k-1}\| \|e^{k-1}\|^2_{L^4(\Omega)} \le c_S \|e^{k-1}\|(\|e^{k-1}\|^2 + \|\nabla e^{k-1}\|^2), \tag{6.3}$$

we find by summation of (6.2) and (6.3) over $k = 1, 2, \ldots, L$ that we are in the situation of Lemma 6.2 with

$$A = \varepsilon^{-2}\tau \sum_{k=1}^{K} \|\mathscr{C}_\tau(t_k)\|^2_{H^1(\Omega)'}, \; E = \exp\Big(2\tau \sum_{k=1}^{K} \mu_\lambda^k\Big), \; B = 48\varepsilon^{-4} c_S,$$

and $\beta = 1/2$. Therefore,

$$\sup_{k=0,\ldots,K} \|e^k\|^2 + \varepsilon^2 \tau \sum_{k=1}^{K} \|\nabla e^k\|^2 \le 8AE,$$

provided that $8AE \le (8B(1+T)E)^{-2}$. Since according to Proposition 6.5 we have $A \le c\tau^2\varepsilon^{-6}$, this is satisfied if $c_B \tau^2 \varepsilon^{-6} E \le (8B(1+T)E)^{-2}$. With the assumed bound for the discrete integral of λ_{AC}^+, we deduce that

$$E \le \exp(8T) \exp\Big(4\tau \sum_{k=1}^{K} \lambda_{AC}^+(t_k)\Big) \le c_E \varepsilon^{-4\kappa}.$$

Therefore, the condition $\tau^2 \le c\varepsilon^{14}\varepsilon^{12\kappa}$ implies the assertion. □

Remarks 6.6 (i) If $u_{tt} \in L^2([0,T]; L^2(\Omega))$, then the bound for A in the proof can be improved and the conditions of the theorem can be weakened.
(ii) An a priori error analysis for a fully discrete approximation follows the same strategy by decomposing the error $u(t_k) - u_h^k$ as $(u(t_k) - Q_h u(t_k)) + (Q_h u(t_k) - u_h^k)$ with the H^1-projection Q_h, cf. [8].

6.3 Practical Realization

We discuss in this section alternatives to the implicit Euler scheme and include an estimate for the approximation of the principal eigenvalue that is needed to compute the a posteriori error bound.

6.3.1 Time-Stepping Schemes

The implicit Euler scheme requires the solution of a nonlinear system of equations in every time step and is stable under the condition $\tau \leq 2\varepsilon^2$. We consider various semi-implicit approximation schemes defined by approximating the nonlinear term avoiding some of these limitations.

Algorithm 6.1 (*Semi-implicit approximation*) Given $u_h^0 \in \mathscr{S}^1(\mathscr{T}_h)$, $\tau > 0$, and a continuous function $G : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ let the sequence $(u_h^k)_{k=0,\dots,K}$ be defined by

$$(d_t u_h^k, v_h) + (\nabla u_h^k, \nabla v_h) + \varepsilon^{-2}\big(G(u_h^k, u_h^{k-1}), v_h\big) = 0$$

for all $v_h \in \mathscr{S}^1(\mathscr{T}_h)$.

The function G is assumed to provide a *consistent* approximation of the nonlinear function f in the sense that $G(s, s) = f(s)$.

Examples 6.1 (i) The (fully) implicit Euler scheme corresponds to

$$G^{\text{impl}}(u^k, u^{k-1}) = f(u^k).$$

(ii) The choice of

$$G^{\text{expl}}(u^k, u^{k-1}) = f(u^{k-1})$$

realizes an explicit treatment of the nonlinearity.
(iii) Carrying out one iteration of a Newton scheme in every time step of the implicit Euler scheme with initial guess u_h^{k-1} corresponds to the linearization

$$G^{\text{lin}}(u^k, u^{k-1}) = f(u^{k-1}) + f'(u^{k-1})(u^k - u^{k-1}).$$

(iv) A Crank–Nicolson type treatment of the nonlinear term is

$$G^{\text{cn}}(u^k, u^{k-1}) = \begin{cases} \dfrac{F(u^k) - F(u^{k-1})}{u^k - u^{k-1}} & \text{if } u^k \neq u^{k-1}, \\ f(u^k) & \text{if } u^k = u^{k-1}. \end{cases}$$

We have $G^{\text{cn}}(u^k, u^{k-1}) = (1/4)(u^k + u^{k-1})((u^k)^2 + (u^{k-1})^2 - 2)$.
(v) The decomposition $F = F^{cx} + F^{cv}$ of $F(u^{k-1}) = ((u^k)^2 - 1)^2/4$ into a convex part $F^{cx}(u^{k-1}) = ((u^k)^4 + 1)/4$ and a concave part $F^{cv}(u^{k-1}) = -(1/2)(u^{k-1})^2$ leads with the derivatives f^{cx} and f^{cv} of F^{cx} and F^{cv}, respectively, to the definition

$$G^{\text{cxcv}}(u^k, u^{k-1}) = f^{cx}(u^k) + f^{cv}(u^{k-1}).$$

Remarks 6.7 (i) Only the explicit and linearized treatment of the nonlinear term leads to linear systems of equations in every time step. The convex-concave decomposition leads to monotone systems of equations.
(ii) The best compromise for stability and linearity appears to be the linearized treatment of the nonlinear term.
(iii) The decomposition of F into convex and concave parts corresponds to the general concept to treat monotone terms implicitly and anti-monotone terms explicitly.
(iv) Numerical integration simplifies the nonlinearities, i.e., for all $z, y \in \mathcal{N}_h$, we have

$$\big(G(u_h^k, u_h^{k-1})\varphi_z, \varphi_y\big)_h = G(u_h^k(z), u_h^{k-1}(z))\beta_z\delta_{zy}$$

with $\beta_z = \int_\Omega \varphi_z$, so that the corresponding contribution to the system matrix is given by a diagonal matrix.
(v) The numerical schemes have different numerical dissipation properties.

The stability of the different semi-implicit Euler schemes is a consequence of the following proposition. We omit a discussion of the explicit treatment of the nonlinearity since this is experimentally found to be unstable even for $\tau \sim \varepsilon^2$.

Proposition 6.6 (Semi-implicit Euler schemes) *Given $u^k, u^{k-1} \in \mathbb{R}$ and $\tau > 0$, we set $d_t u^k = (u^k - u^{k-1})/\tau$. We have*

$$\begin{aligned} G^{\text{impl}}(u^k, u^{k-1})d_t u^k &\ge d_t F(u^k) - \frac{\tau}{2}|d_t u^k|^2, \\ G^{\text{cn}}(u^k, u^{k-1})d_t u^k &= d_t F(u^k), \\ G^{\text{cxcv}}(u^k, u^{k-1})d_t u^k &\ge d_t F(u^k), \end{aligned}$$

and if $|u^k|, |u^{k-1}| \le 1$, then

$$G^{\text{lin}}(u^k, u^{k-1})d_t u^k \ge d_t F(u^k) - \frac{7\tau}{2}|d_t u^k|^2.$$

In particular, the implicit Euler scheme is stable if $\tau \le 2\varepsilon^2$, the semi-implicit Euler scheme with Crank–Nicolson type treatment of the nonlinear term is unconditionally stable, the semi-implicit Euler scheme with decomposed treatment of the nonlinearity is unconditionally stable, and the semi-implicit Euler scheme with a linearized treatment of the nonlinear term is stable if a discrete maximum principle holds and $\tau \le (2/7)\varepsilon^2$, i.e., under these conditions we have for the solutions of the respective semi-implicit Euler schemes that

$$I_\varepsilon(u_h^L) \le I_\varepsilon(u_h^0)$$

for all $L \ge 0$.

Proof A Taylor expansion shows that for some $\xi \in \mathbb{R}$, we have

$$F(u^{k-1}) = F(u^k) + f(u^k)(u^{k-1} - u^k) + \frac{1}{2} f'(\xi)(u^{k-1} - u^k)^2.$$

Since $f'(\xi) \geq -1$ we deduce after division by τ that

$$f(u^k) d_t u^k = d_t F(u^k) + \frac{\tau}{2} f'(\xi)(d_t u^k)^2 \geq d_t F(u^k) - \frac{\tau}{2} |d_t u^k|^2$$

and this implies the bound for G^{impl}. Assuming that $|u^k|, |u^{k-1}| \leq 1$, a similar argument with $f''(s) = 6s$ shows with some $\zeta \in [-1, 1]$ that

$$\begin{aligned}&\big(f(u^{k-1}) + f'(u^{k-1})(u^k - u^{k-1})\big)(u^k - u^{k-1})\\ &= f(u^k)(u^k - u^{k-1}) - \frac{1}{2} f''(\zeta)(u^k - u^{k-1})^3 \geq f(u^k)(u^k - u^{k-1}) - 6(u^k - u^{k-1})^2\end{aligned}$$

and with the previous estimate we deduce that

$$G^{\text{lin}}(u^k, u^{k-1}) d_t u^k \geq d_t F(u^k) - \frac{7\tau}{2} |d_t u^k|^2.$$

If $d_t u^k \neq 0$, then

$$G^{\text{cn}}(u^k, u^{k-1})(u^k - u^{k-1}) = F(u^k) - F(u^{k-1}) = \tau d_t F(u^k),$$

and if $d_t u^k = 0$, then $G^{\text{cn}}(u^k, u^{k-1}) d_t u^k = 0 = \tau d_t F(u^k)$ which implies the asserted identity for G^{cn}. For the convex function F^{cx} and its derivative f^{cx}, we have

$$f^{cx}(u^k)(u^{k-1} - u^k) + F^{cx}(u^k) \leq F^{cx}(u^{k-1}).$$

Analogously, for the convex function $-F^{cv}$ and its derivative $-f^{cv}$, we have

$$-f^{cv}(u^{k-1})(u^k - u^{k-1}) - F^{cv}(u^{k-1}) \leq -F^{cv}(u^k).$$

The combination of the two estimates proves that

$$G^{\text{cxcv}}(u^k, u^{k-1}) d_t u^k = f^{cx}(u^k) d_t u^k + f^{cv}(u^{k-1}) d_t u^k \geq d_t F^{cx}(u^k) + d_t F^{cv}(u^k).$$

The stability of the related schemes now follows from the choice of $v_h = d_t u_h^k$ in the semi-implicit Euler scheme, i.e.,

$$\|d_t u_h^k\|^2 + \frac{d_t}{2} \|\nabla u_h^k\|^2 + \frac{\tau}{2} \|\nabla d_t u_h^k\|^2 + \varepsilon^{-2} \big(G(u_h^k, u_h^{k-1}), d_t u_h^k\big) = 0,$$

together with a summation over $k = 1, 2, \dots, L$, and the corresponding lower bounds for $G(u_h^k, u_h^{k-1})$. □

6.3.2 Computation of the Eigenvalue

The a posteriori error estimate of Theorem 6.5 requires a lower bound for the principal eigenvalue of the linearized Allen–Cahn operator with respect to the approximate solution, i.e., a function λ_{AC}^h such that

$$-\lambda_{\mathrm{AC}}^h(t) \le \inf_{v \in H^1(\Omega)\setminus\{0\}} \frac{\|\nabla v\|^2 + \varepsilon^{-2}(f'(u_{h,\tau}(t))v, v)}{\|v\|^2}.$$

To approximate the infimum on the right-hand side, we replace the space $H^1(\Omega)$ by $\mathcal{S}^1(\mathcal{T}_h)$. We fix a time t in the following and let $-\Lambda \in \mathbb{R}$ be the infimum at time t, i.e., there exists $w \in H^1(\Omega)$ with $\|w\| = 1$ and

$$-\Lambda(w, v) = (\nabla w, \nabla v) + \varepsilon^{-2}(p_h w, v)$$

for all $v \in H^1(\Omega)$ and with $p_h = f'(u_{h,\tau}(t))$.

Proposition 6.7 (Eigenvalue approximation) *Let $(\Lambda_h, w_h) \in \mathbb{R} \times \mathcal{S}^1(\mathcal{T}_h)$ be such that*

$$-\Lambda_h(w_h, v_h) = (\nabla w_h, \nabla v_h) + \varepsilon^{-2}(p_h w_h, v_h)$$

for all $v_h \in \mathcal{S}^1(\mathcal{T}_h)$. Assume that the Laplace operator with homogeneous Neumann boundary conditions is H^2-regular in Ω in the sense that $\|D^2 v\| \le c_\Delta \|\Delta v\|$ for all $v \in H^2(\Omega)$ with $\partial_n v = 0$ on $\partial\Omega$ and suppose that $\|p_h\|_{L^\infty(\Omega)} \le c_0$. Then there exists $c_1 > 0$ such that if $h \le c_1 \varepsilon$, we have

$$0 \le \Lambda - \Lambda_h \le c\varepsilon^{-4}h^2.$$

Proof In the following we occasionally replace the function p_h by $q_h = p_h + \|p_h\|_{L^\infty(\Omega)}$ which corresponds to a shift of $-\Lambda$ and $-\Lambda_h$ by $\|p_h\|_{L^\infty(\Omega)}$ but allows us to use $q_h \ge 0$. The fact that $\mathcal{S}^1(\mathcal{T}_h) \subset H^1(\Omega)$ implies that we have $-\Lambda \le -\Lambda_h$. Since w_h is minimal for $v_h \mapsto \|\nabla v_h\|^2 + \varepsilon^{-2}(p_h v_h, v_h)$ among functions $v_h \in \mathcal{S}^1(\mathcal{T}_h)$ with $\|v_h\| = 1$ with minimum $-\Lambda_h$ and since $-\Lambda = \|\nabla w\|^2 + \varepsilon^{-2}(p_h w, w)$, we have

$$\begin{aligned} 0 \le \Lambda - \Lambda_h &\le -(\nabla w, \nabla w) - \varepsilon^{-2}(q_h w, w) + \|\nabla v_h\|^2 + \varepsilon^{-2}(q_h v_h, v_h) \\ &\le 2(\nabla v_h, \nabla[v_h - w]) + 2\varepsilon^{-2}(q_h v_h, v_h - w). \end{aligned}$$

We note $-\Lambda \le \varepsilon^{-2}\|p_h\|_{L^\infty(\Omega)}$ and conclude with $-\Delta w = -\Lambda w - \varepsilon^{-2} p_h w$ that

$$\|\nabla w\| \le c\varepsilon^{-1}, \quad \|D^2 w\| \le c\|\Delta w\| \le c\varepsilon^{-2}.$$

We incorporate the H^1-projection $Q_h w \in \mathscr{S}^1(\mathscr{T}_h)$ defined by

$$(\nabla Q_h w, \nabla y_h) + (Q_h w, y_h) = (\nabla w, \nabla y_h) + (w, y_h)$$

for all $y_h \in \mathscr{S}^1(\mathscr{T}_h)$ which satisfies the estimates

$$h\|w - Q_h w\| + \|\nabla(w - Q_h w)\| \le ch^2\|D^2 w\|.$$

We suppose that $h \le c\varepsilon$ is such that

$$\big|1 - \|Q_h w\|\big| \le \|w - Q_h w\| \le ch^2\varepsilon^{-2} \le \frac{1}{2}.$$

Choosing $v_h = Q_h w/\|Q_h w\|$ and noting

$$\|\nabla Q_h w\| + \|Q_h w\| \le \|\nabla w\| + \|w\| \le c\varepsilon^{-1}$$

we find that

$$\begin{aligned}(\nabla v_h, \nabla[v_h - w]) &= \|Q_h w\|^{-2}\big((\nabla Q_h w, \nabla[Q_h w - w]) + (\nabla Q_h w, \nabla[w - \|Q_h w\| w])\big)\\ &= \|Q_h w\|^{-2}\big((Q_h w, Q_h w - w) + (1 - \|Q_h w\|)(\nabla Q_h w, \nabla w)\big)\\ &\le ch^2\varepsilon^{-2}(1 + \varepsilon^{-2}).\end{aligned}$$

Analogously, we have

$$\begin{aligned}(q_h v_h, v_h - w) &= \|Q_h w\|^{-2}\big((Q_h w, Q_h w - w) + (Q_h w, w - \|Q_h w\| w)\big)\\ &= \|Q_h w\|^{-2}\big((Q_h w, Q_h w - w) + (1 - \|Q_h w\|)(Q_h w, w)\big)\\ &\le ch^2\varepsilon^{-2}.\end{aligned}$$

A combination of the estimates implies the asserted error bound. □

The discrete eigenvalue problem can be recast as the problem of finding a vector $W \in \mathbb{R}^L$ with $W^\top m W = 1$ and

$$(-\Lambda + c_{\text{shift}}) m W = (s + \varepsilon^{-2} m_p + c_{\text{shift}} m) W = Y W$$

with the mass matrix m, the stiffness matrix s, the weighted mass matrix m_p, and an arbitrary constant c_{shift}. For $c_{\text{shift}} = \varepsilon^{-2}\|p_h\|_{L^\infty(\Omega)} + 1$, we have that the symmetric matrices m and $Y = s + \varepsilon^{-2} m_p + c_{\mathrm{shift}} m$ are positive definite, and we may use the following vector iteration with Rayleigh-quotient approximation to approximate Λ.

Algorithm 6.2 (*Vector iteration*) Given $W_0 \in \mathbb{R}^L$ such that $W_0^\top m W_0 = 1$, compute the sequence Λ^j, $j = 0, 1, 2, \dots$ via $\Lambda^0 = (W^0)^\top Y W^0$ and

$$\widetilde{W}^{j+1} = Y^{-1}(m W^j), \quad W^{j+1} = \frac{\widetilde{W}^{j+1}}{\big((\widetilde{W}^{j+1})^\top m \widetilde{W}^{j+1}\big)^{1/2}}$$

and

$$-\Lambda^{j+1} + c_{\mathrm{shift}} = (W^{j+1})^\top Y W^{j+1}.$$

Stop the iteration if $|\Lambda^{j+1} - \Lambda^j| \le \varepsilon_{\mathrm{stop}}$.

Remark 6.8 The iteration converges to the smallest eigenvalue provided that the initial vector W_0 is not orthogonal to the corresponding eigenspace.

6.3.3 Implementation

The MATLAB code shown in Fig. 6.6 realizes the semi-implicit Euler scheme with linearized treatment of the nonlinear term and computes the principal eigenvalue defined by the approximate solution in every time step. We used the discrete inner product $(\cdot, \cdot)_h$ to simplify the computation of some matrices, i.e., we use the formulations

$$\begin{aligned}(d_t u_h^k, v_h) + (\nabla u_h^k, \nabla v_h) + \varepsilon^{-2}(f'(u_h^{k-1}) u_h^k, v_h)_h \\ = -\varepsilon^{-2}(f(u_h^{k-1}), v_h)_h + \varepsilon^{-2}(f'(u_h^{k-1}) u_h^{k-1}, v_h)_h\end{aligned}$$

and

$$-\lambda_{\mathrm{AC}}^h(t_k)(w_h, v_h) = (\nabla w_h, \nabla v_h) + \varepsilon^{-2}(f'(u_h^k) w_h, v_h)_h$$

for all $v_h \in \mathscr{S}^1(\mathscr{T}_h)$ to find $u_h^k \in \mathscr{S}^1(\mathscr{T}_h)$ and an approximation of the eigenpair $(-\lambda_{\mathrm{AC}}^h(t_k), w_h)$.

```
function ac_linearized_euler(d,red)
[c4n,n4e,Db,Nb] = triang_cube(d);
c4n = 2*(c4n-1/2);
for j = 1:red
    [c4n,n4e,Db,Nb,¬,¬] = red_refine(c4n,n4e,Db,Nb);
end
nC = size(c4n,1);
T = 1;
eps = 2^(-4); tau = (2/3)*eps^2;
K = ceil(T/tau);
u = u_0(c4n,eps);
[s,m,m_lumped] = fe_matrices(c4n,n4e);
w_init = rand(nC,1)-.5;
lambda = zeros(K,1);
for k = 1:K
    b_nonlin = -eps^(-2)*m_lumped*f(u)...
        +eps^(-2)*m_lumped*(df(u).*u);
    m_nonlin = eps^(-2)*m_lumped*diag(df(u));
    b = tau^(-1)*m*u+b_nonlin;
    X = tau^(-1)*m+s+m_nonlin;
    u = X\b;
    c_shift = abs(min(df(u)))+1;
    Y = s+eps^(-2)*m_lumped*spdiags(df(u)+c_shift,0,nC,nC);
    [neg_lambda_shift,w] = vector_iteration(Y,m,w_init);
    lambda(k) = -neg_lambda_shift+eps^(-2)*c_shift;
    figure(1); show_p1(c4n,n4e,Db,Nb,u); axis square;
    figure(2); plot(tau*(1:k),lambda(1:k)); drawnow;
    w_init = w;
end

function val = f(u)
val = u.^3-u;
function val = df(u)
val = 3*u.^2-1;

function val = u_0(x,eps)
dist = sqrt(min((x(:,1)-.3).^2,(x(:,1)+.3).^2)+x(:,2).^2)-.35;
val = -tanh(dist/(sqrt(2)*eps));

function [mu,w] = vector_iteration(Y,m,w)
mu = 0; mu_old = 0;
diff_mu = 1; eps_stop = 1E-01;
while abs(diff_mu) > eps_stop
    w = Y\(m*w);
    w = w/sqrt(w'*m*w);
    mu = w'*Y*w;
    diff_mu = mu-mu_old;
    mu_old = mu;
end
```

Fig. 6.6 Implementation of the linearized implicit Euler scheme with numerical integration for the Allen–Cahn equation and computation of the eigenvalue in each time step

References

1. Alikakos, N.D., Fusco, G.: Slow dynamics for the Cahn-Hilliard equation in higher space dimensions. I. Spectral estimates. Commun. Partial Differ Equ **19**(9–10), 1397–1447 (1994). http://dx.doi.org/10.1080/03605309408821059
2. Barrett, J.W., Blowey, J.F.: An error bound for the finite element approximation of the Cahn-Hilliard equation with logarithmic free energy. Numer. Math. **72**(1), 1–20 (1995). http://dx.doi.org/10.1007/s002110050157
3. Bartels, S., Müller, R., Ortner, C.: Robust a priori and a posteriori error analysis for the approximation of Allen-Cahn and Ginzburg-Landau equations past topological changes. SIAM J. Numer. Anal. **49**(1), 110–134 (2011). http://dx.doi.org/10.1137/090751530
4. Chen, X.: Spectrum for the Allen-Cahn, Cahn-Hilliard, and phase-field equations for generic interfaces. Commun. Partial Differ Equ **19**(7–8), 1371–1395 (1994). http://dx.doi.org/10.1080/03605309408821057
5. Deckelnick, K., Dziuk, G., Elliott, C.M.: Computation of geometric partial differential equations and mean curvature flow. Acta Numer. **14**, 139–232 (2005). http://dx.doi.org/10.1017/S0962492904000224
6. Elliott, C.M., French, D.A.: Numerical studies of the Cahn-Hilliard equation for phase separation. IMA J. Appl. Math. **38**(2), 97–128 (1987). http://dx.doi.org/10.1093/imamat/38.2.97
7. Emmerich, H.: The Diffuse Interface Approach in Materials Science. Lecture Notes in Physics, vol. M 73. Springer, Berlin (2003)
8. Feng, X., Prohl, A.: Numerical analysis of the Allen-Cahn equation and approximation for mean curvature flows. Numer. Math. **94**(1), 33–65 (2003). http://dx.doi.org/10.1007/s00211-002-0413-1
9. Kessler, D., Nochetto, R.H., Schmidt, A.: A posteriori error control for the Allen-Cahn problem: circumventing Gronwall's inequality. M2AN Math. Model. Numer. Anal. **38**(1), 129–142 (2004). http://dx.doi.org/10.1051/m2an:2004006
10. Nochetto, R.H., Verdi, C.: Convergence past singularities for a fully discrete approximation of curvature-driven interfaces. SIAM J. Numer. Anal. **34**(2), 490–512 (1997). http://dx.doi.org/10.1137/S0036142994269526
11. Penrose, O., Fife, P.C.: Thermodynamically consistent models of phase-field type for the kinetics of phase transitions. Phys. D **43**(1), 44–62 (1990). http://dx.doi.org/10.1016/0167-2789(90)90015-H

Chapter 7
Harmonic Maps

7.1 Analytical Properties

Harmonic maps are stationary points of the Dirichlet energy in the set of vector fields that attain their values in a given target manifold, e.g., the unit sphere. Related problems arise in various applications and the problem of computing harmonic maps serves as a mathematical model problem for constrained minimization problems on infinite-dimensional spaces. We will consider the case of computing harmonic maps into the unit sphere $S^{m-1} = \{s \in \mathbb{R}^m : |s| = 1\}$, i.e., unit-length vector fields, but notice that a large class of target manifolds can be treated with the same ideas. We thus aim at approximating minimizers $u \in \mathscr{A}$ for

$$I(u) = \frac{1}{2} \int_\Omega |\nabla u|^2 \,\mathrm{d}x$$

with the set of admissible vector fields

$$\mathscr{A} = \{v \in H^1(\Omega; \mathbb{R}^m) : |v(x)| = 1 \text{ for a.e. } x \in \Omega,\ v|_{\Gamma_{\mathrm{D}}} = u_{\mathrm{D}}\}.$$

The function $u_{\mathrm{D}} \in L^2(\Gamma_{\mathrm{D}}; \mathbb{R}^m)$ on the nonempty set $\Gamma_{\mathrm{D}} \subset \partial\Omega$ is assumed to admit an extension $\widetilde{u}_{\mathrm{D}} \in H^1(\Omega; \mathbb{R}^m)$ with $|\widetilde{u}_{\mathrm{D}}(x)| = 1$ for almost every $x \in \Omega$. We briefly summarize the main properties of harmonic maps and refer the reader to the textbooks [9, 12] for more details.

7.1.1 Existence and Nonuniqueness

The existence of minimizers is established by the direct method in the calculus of variations.

S. Bartels, *Numerical Methods for Nonlinear Partial Differential Equations*,
Springer Series in Computational Mathematics 47,
DOI 10.1007/978-3-319-13797-1_7

Theorem 7.1 (Existence) *There exists a minimizer $u \in \mathscr{A}$.*

Proof Since u_{D} admits an extension to a unit-length vector field field $\widetilde{u}_{\mathrm{D}} \in \mathscr{A}$ there exists an infimizing sequence $(u_j)_{j\in\mathbb{N}} \subset \mathscr{A}$ with $\lim_{j\to\infty} I(u_j) = \inf_{v\in\mathscr{A}} I(v)$. Since $u_j - \widetilde{u}_{\mathrm{D}} \in H^1_{\mathrm{D}}(\Omega;\mathbb{R}^m)$, we have that $(u_j)_{j\in\mathbb{N}}$ is bounded in $H^1(\Omega;\mathbb{R}^m)$. A subsequence converges weakly to a vector field $u \in H^1(\Omega;\mathbb{R}^m)$ with $u|_{\Gamma_{\mathrm{D}}} = u_{\mathrm{D}}$. To show that $u \in \mathscr{A}$ we notice that the subsequence converges strongly in $L^2(\Omega;\mathbb{R}^m)$, and hence there exists a further subsequence that converges pointwise almost everywhere to u. Therefore, $|u| = 1$ almost everywhere in Ω, i.e., $u \in \mathscr{A}$. The weak lower semicontinuity of I implies that u is a minimizer. □

Remark 7.1 The proof shows that the set $\mathscr{A}$ is weakly closed.

The essential condition that $\mathscr{A} \neq \emptyset$ may be difficult to verify in practice even if $u_{\mathrm{D}} \in L^2(\Gamma_{\mathrm{D}};\mathbb{R}^m)$ is smooth and satisfies $|u_{\mathrm{D}}(x)| = 1$ for almost every $x \in \partial\Omega$.

Example 7.1 (Nonexistence) For $\Omega = B_1(0) \subset \mathbb{R}^2$ and $u_{\mathrm{D}}(x) = x$ there is no function $\widetilde{u}_{\mathrm{D}} \in H^1(\Omega;\mathbb{R}^2)$ with $\widetilde{u}_{\mathrm{D}}|_{\partial\Omega} = u_{\mathrm{D}}$ and $|\widetilde{u}_{\mathrm{D}}(x)| = 1$ for almost every $x \in \Omega$. This is a consequence of the Hopf–Poincaré formula and Brouwer's fixed point theorem.

Due to the invariance of the Dirichlet energy under rotations, we cannot expect harmonic maps to be unique.

Example 7.2 (Nonuniqueness) Let $\Omega = (0,1)$, $\Gamma_{\mathrm{D}} = \partial\Omega = \{0,1\}$, $m = 3$, and let $u : (0,1) \to S^2$ be minimal for

$$I(u) = \frac{1}{2}\int_0^1 |u'|^2 \,\mathrm{d}x$$

in the set of functions $v \in \mathscr{A}$ with $v(0) = e$ and $v(1) = -e$ for some $e \in S^2$. Then for every rotation $Q \in SO(3) = \{R \in \mathbb{R}^{3\times 3} : R^{\top}R = I_3,\ \det R = 1\}$ with $Qe = e$, we have that $\widetilde{u} = Qu$ is another minimizer. The harmonic maps $u_1(x) = [\cos(\pi x), 0, \sin(\pi x)]^{\top}$, $x \in (0,1)$, and $u_2 = Qu_1$, where $Q \in \mathbb{R}^{3\times 3}$ realizes a rotation by π about the first coordinate axis, with identical Dirichlet energy are shown in Fig. 7.1.

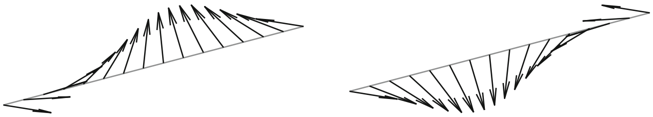


Fig. 7.1 Two harmonic maps on $\Omega = (0,1)$ with the same boundary values and identical Dirichlet energy; the length of the *arrows* is scaled for graphical purposes

Remarks 7.2 (i) Harmonic maps can be approximated by penalizing the pointwise constraint, e.g., considering for $\varepsilon > 0$ the Ginzburg–Landau regularization

$$I_\varepsilon(u) = \frac{1}{2}\int_\Omega |\nabla u|^2 \,\mathrm{d}x + \frac{\varepsilon^{-2}}{4}\int_\Omega (|u|^2 - 1)^2 \,\mathrm{d}x$$

and investigating the limiting behavior of minimizers $(u_\varepsilon)_{\varepsilon>0}$ as $\varepsilon \to 0$.
(ii) Formally, a harmonic map u and the Lagrange multiplier λ associated to the length constraint define a saddle-point for the functional

$$L(u, \lambda) = \frac{1}{2}\int_\Omega |\nabla u|^2 \,\mathrm{d}x + \int_\Omega \lambda(|u|^2 - 1) \,\mathrm{d}x.$$

7.1.2 Euler–Lagrange Equations and Nonregularity

The Euler–Lagrange equations define a nonlinear partial differential equation.

Theorem 7.2 (Euler–Lagrange equations) *Let $u \in \mathscr{A}$ be stationary for the Dirichlet energy. Then we have*

$$(\nabla u, \nabla w) = (|\nabla u|^2 u, w)$$

for all $w \in H^1_{\mathrm{D}}(\Omega; \mathbb{R}^m) \cap L^\infty(\Omega; \mathbb{R}^m)$.

Proof Let $w \in H^1(\Omega; \mathbb{R}^m) \cap L^\infty(\Omega; \mathbb{R}^m)$ and $\varepsilon > 0$ be such that $\varepsilon \|w\|_{L^\infty(\Omega)} \le 1/2$. We then have that $|u(x) + rw(x)| \ge 1/2$ for almost every $x \in \Omega$ and every $r \in \mathbb{R}$ with $|r| \le \varepsilon$. It follows that the map

$$u^r(x) = \frac{u(x) + rw(x)}{|u(x) + rw(x)|}$$

belongs to $H^1(\Omega; \mathbb{R}^m)$ and satisfies $|u^r| = 1$ in Ω and $u^r|_{\Gamma_{\mathrm{D}}} = u_{\mathrm{D}}$. Since $u^0 = u$, we have that the function $t \mapsto I(u^r)$ is minimal at $r = 0$. We note that

$$\frac{d}{dr}\Big|_{r=0} u^r = w - u(u \cdot w).$$

A differentiation shows that

$$0 = \frac{d}{dr}\Big|_{r=0} I(u^r) = \sum_{\ell=1}^{d} \int_\Omega \partial_\ell u \cdot \partial_\ell \big[w - u(u \cdot w)\big] \,\mathrm{d}x$$

and the orthogonality $(\partial_\ell u) \cdot u = 0$ for $\ell = 1, 2, \ldots, d$ implies the assertion. □

Definition 7.1 Solutions $u \in \mathscr{A}$ of the Euler–Lagrange equation are called *harmonic maps* (*into the sphere*).

Remark 7.3 The function $\lambda = |\nabla u|^2 \in L^1(\Omega)$ is the Lagrange multiplier associated to the pointwise constraint $|u(x)|^2 = 1$.

Solutions of the Euler–Lagrange equations are in general neither energy minimizing nor regular.

Example 7.3 (Nonregularity) Let $\Omega = (-1, 1)^3$ and $u_{\mathrm{D}}(x) = x/|x|$ for $x \in \Gamma_{\mathrm{D}} = \partial\Omega$. Then $u(x) = x/|x|$ for $x \in \Omega$ satisfies $u \in \mathscr{A}$ and is a harmonic map. Moreover u is minimal for I in the set of vector fields in $\mathscr{A}$.

Remarks 7.4 (i) For $d = 2$, harmonic maps are smooth.
(ii) If $d = 3$, then energy minimizing harmonic maps u are partially regular in the sense that u is smooth in $\Omega \setminus S$ for a set S with $\mathscr{H}^1(S) = 0$, e.g., a set of points. Harmonic maps that are not globally energy minimizing can be discontinuous everywhere.

7.1.3 Compactness

The lack of uniqueness and regularity of harmonic maps makes it difficult to quantify stability properties. The weaker concept of compactness shows that accumulation points of (almost) harmonic maps are again harmonic maps, i.e., that bounded subsets of the set of harmonic maps are weakly compact. The key to this property is the following equivalent characterization of harmonic maps. We restrict ourselves to the case $m = 3$ for ease of presentation.

Lemma 7.1 (Equivalent characterization) *Let $m = 3$. The function $u \in \mathscr{A}$ is a harmonic map if and only if*

$$(\nabla u, \nabla[u \times \phi]) = \sum_{\ell=1}^{d} (\partial_\ell u, u \times \partial_\ell \phi) = 0$$

for all $\phi \in H^1_{\mathrm{D}}(\Omega; \mathbb{R}^3) \cap L^\infty(\Omega; \mathbb{R}^3)$. This is the case if and only if

$$(\nabla u, \nabla w) = 0$$

for all $w \in H^1_{\mathrm{D}}(\Omega; \mathbb{R}^3)$ satisfying $u \cdot w = 0$ almost everywhere in Ω.

Proof (i) Let $u \in \mathscr{A}$ be a harmonic map. Then the choice $w = u \times \phi$ in the Euler–Lagrange equations, the fact that $u \cdot (u \times \phi) = 0$, and the identity

$$(\partial_\ell u, \partial_\ell [u \times \phi]) = (\partial_\ell u, u \times \partial_\ell \phi)$$

for $\ell = 1, 2, \dots, d$ imply the first characterization. The second one is an immediate consequence of the Euler–Lagrange equations if $w \in H^1_{\mathrm{D}}(\Omega; \mathbb{R}^3) \cap L^\infty(\Omega; \mathbb{R}^3)$ with $w \cdot u = 0$ in Ω. A truncation argument shows that this is satisfied for all $w \in H^1_{\mathrm{D}}(\Omega; \mathbb{R}^3)$ with $w \cdot u = 0$ almost everywhere in Ω.
(ii) Assume that the first equation of the lemma is satisfied and let $w \in H^1_{\mathrm{D}}(\Omega; \mathbb{R}^3) \cap L^\infty(\Omega; \mathbb{R}^3)$. For $\phi = u \times w$ we have, due to the formula $a \times (b \times c) = b(a \cdot c) - c(a \cdot b)$ that

$$u \times \phi = u \times (u \times w) = (u \cdot w)u - |u|^2 w = (u \cdot w)u - w.$$

Moreover, we have for $\ell = 1, 2, \dots, d$ that

$$\partial_\ell[(u \cdot w)u] = (\partial_\ell u \cdot w)u + (u \cdot \partial_\ell w)u + (u \cdot w)\partial_\ell u.$$

With $\partial_\ell u \cdot u = 0$ this implies that

$$0 = \sum_{\ell=1}^{d} \big[(\partial_\ell u, (u \cdot w)\partial_\ell u) - (\partial_\ell u, \partial_\ell w)\big] = (|\nabla u|^2 u, w) - (\nabla u, \nabla w)$$

which proves that u is a harmonic map.
(iii) Suppose that the second characterization is satisfied and let $\phi \in H^1_{\mathrm{D}}(\Omega; \mathbb{R}^3) \cap L^\infty(\Omega; \mathbb{R}^3)$. The function $w = u \times \phi$ satisfies $u \cdot w = 0$ so that the first characterization holds. □

Remark 7.5 The condition that $(\nabla u, \nabla w) = 0$ for all $w \in H^1_{\mathrm{D}}(\Omega; \mathbb{R}^3)$ satisfying $u \cdot w = 0$ shows that u is stationary with respect to tangential perturbations.

The equivalent characterizations imply the following weak compactness result which will serve as a guideline to prove convergence of numerical approximations.

Theorem 7.3 (Weak compactness) *Let $(\mathscr{R}_j)_{j \in \mathbb{N}} \subset H^1_{\mathrm{D}}(\Omega; \mathbb{R}^3)'$ be a sequence of functionals with $\|\mathscr{R}_j\|_{H^1_{\mathrm{D}}(\Omega)'} \to 0$ as $j \to \infty$, and assume that $(u_j)_{j \in \mathbb{N}} \subset \mathscr{A}$ is such that*

$$(\nabla u_j, \nabla w) = (|\nabla u_j|^2 u_j, w) + \mathscr{R}_j(w)$$

for every $j \in \mathbb{N}$ and all $w \in H^1_{\mathrm{D}}(\Omega; \mathbb{R}^3) \cap L^\infty(\Omega; \mathbb{R}^3)$. If $u \in H^1(\Omega; \mathbb{R}^3)$ is such that $u_j \rightharpoonup u$ in $H^1(\Omega; \mathbb{R}^3)$ as $j \to \infty$, then we have $u \in \mathscr{A}$ and u is a harmonic map.

Proof The weak closedness of $\mathscr{A}$ implies that $u \in \mathscr{A}$. For every $\phi \in H^1_{\mathrm{D}}(\Omega; \mathbb{R}^3) \cap C^\infty(\overline{\Omega}; \mathbb{R}^3)$ and $j \in \mathbb{N}$, the choice of $w = u_j \times \phi$ yields, using $\partial_\ell u_j \cdot (\partial_\ell u_j \times \phi) = 0$,

$$\sum_{\ell=1}^{d} (\partial_\ell u_j, u_j \times \partial_\ell \phi) = \mathscr{R}_j(u_j \times \phi).$$

Since $u_j \to u$ in $L^2(\Omega; \mathbb{R}^3)$ and $\partial_\ell u_j \rightharpoonup \partial_\ell u$ in $L^2(\Omega; \mathbb{R}^3)$, we have

$$(\partial_\ell u_j, u_j \times \partial_\ell \phi) = (\partial_\ell u_j, u \times \partial_\ell \phi) + (\partial_\ell u_j, [u_j - u] \times \partial_\ell \phi)$$
$$\to (\partial_\ell u, u \times \partial_\ell \phi)$$

as $j \to \infty$ for $\ell = 1, 2, \ldots, d$. Employing $\mathscr{R}_j \to 0$ in $H^1_{\mathrm{D}}(\Omega; \mathbb{R}^3)'$ and that $u_j \times \phi$ is bounded in $H^1_{\mathrm{D}}(\Omega; \mathbb{R}^3)$, we also have

$$\mathscr{R}_j(u_j \times \phi) \to 0$$

as $j \to \infty$. Altogether we find that u satisfies

$$\sum_{\ell=1}^{d} (\partial_\ell u, u \times \partial_\ell \phi) = 0$$

for all $\phi \in H^1_{\mathrm{D}}(\Omega; \mathbb{R}^3) \cap C^\infty(\overline{\Omega}; \mathbb{R}^3)$. A density argument shows that this identity holds for all $\phi \in H^1_{\mathrm{D}}(\Omega; \mathbb{R}^3) \cap L^\infty(\Omega; \mathbb{R}^3)$ so that Lemma 7.1 implies that u is a harmonic map. □

Remarks 7.6 (i) The equivalent characterization of harmonic maps involving the cross product allowed us to use that the product of a weakly and a strongly convergent sequence is weakly convergent. We remark that the identification of the limit of the square of a weakly convergent sequence is difficult in general and a passage to a limit in the Euler–Lagrange equations for harmonic maps does not imply that the limit is a harmonic map.
(ii) While the existence of harmonic maps into general target manifolds other than the unit sphere can be established analogously, related compactness results are false in general. For $d = 2$ and sufficiently smooth target manifolds, regularity and compactness can be proved, cf. [11].

7.1.4 Harmonic Map Heat Flow

The harmonic map heat flow is the L^2-gradient flow of the Dirichlet energy subject to the unit length constraint and is given by

$$\partial_t u - \Delta u = |\nabla u|^2 u, \;\; |u(t, \cdot)| = 1, \;\; u(0) = u_0, \;\; u|_{\Gamma_{\mathrm{D}}} = u_{\mathrm{D}}, \;\; \partial_n u|_{\Gamma_{\mathrm{N}}} = 0$$

for almost every $t \in [0, T]$. To avoid very irregular solutions, it is important to construct solutions that satisfy an energy law.

Theorem 7.4 (Existence) *Given $u_0 \in H^1(\Omega; \mathbb{R}^m)$ with $|u_0(x)| = 1$ for almost every $x \in \Omega$, there exists $u \in H^1([0, T]; L^2(\Omega; \mathbb{R}^m)) \cap L^\infty([0, T]; H^1(\Omega; \mathbb{R}^m))$ such that $|u(t, x)| = 1$ for almost every $(t, x) \in [0, T] \times \Omega$, $u(0) = u_0$,*

$$(\partial_t u, w) + (\nabla u, \nabla w) = (|\nabla u|^2 u, w)$$

for almost every $t \in [0, T]$ and all $w \in H^1_{\mathrm{D}}(\Omega; \mathbb{R}^m) \cap L^\infty(\Omega; \mathbb{R}^m)$, and

$$I(u(T')) + \int_0^{T'} \|\partial_t u\|^2 \, \mathrm{d}t \le I(u_0)$$

for almost every $T' \in [0, T]$.

Proof The result follows from the convergence of numerical approximations proved below. □

Remark 7.7 Uniqueness of solutions is known within the class of energy decreasing solutions if $d = 2$.

Solutions of the harmonic map heat flow can develop singularities in finite time.

Example 7.4 (Finite-time blowup [8]) Let $\Omega = B_1(0) \subset \mathbb{R}^2$, $\Gamma_{\mathrm{D}} = \partial\Omega$, and $u_{\mathrm{D}} = u_0|_{\Gamma_{\mathrm{D}}}$ for u_0 defined for $b > 0$ by

$$u_0(x) = \frac{1}{|x|}\big(x_1 \sin h(|x|), x_2 \sin h(|x|), |x| \cos h(|x|)\big)$$

for $x \in \Omega \setminus \{0\}$ and $h(r) = br^2$. If and only if $b \ge \pi$, the corresponding solution of the harmonic map heat flow is singular in the sense that there exists $T_c > 0$ with $\lim_{t \to T_c} \|\nabla u(t)\|_{L^\infty(\Omega)} = \infty$.

7.2 Numerical Approximation

We discuss in this section the approximation of harmonic maps and employ arguments from [1, 3, 5, 6, 10].

7.2.1 Discrete Harmonic Maps

It is straightforward to verify that the only polynomial vector fields that are pointwise of unit length are constant vector fields. Therefore, the constraint cannot be imposed almost everywhere on polynomial finite element functions. The following proposition shows that it is sufficient to impose the constraint at the nodes of a triangulation, cf. Fig. 7.2.

Proposition 7.1 (Nodal constraint) *Let $(\mathscr{T}_h)_{h>0}$ be a family of regular triangulations of $\Omega \subset \mathbb{R}^d$ and let $(u_h)_{h>0} \subset H^1(\Omega; \mathbb{R}^m)$ be such that $u_h \in \mathscr{S}^1(\mathscr{T}_h)^m$ and $|u_h(z)| = 1$ for all $z \in \mathscr{N}_h$ and every $h > 0$. If $u_h \rightharpoonup u$ in $H^1(\Omega; \mathbb{R}^m)$ for some $u \in H^1(\Omega; \mathbb{R}^m)$, then we have $|u(x)| = 1$ for almost every $x \in \Omega$.*

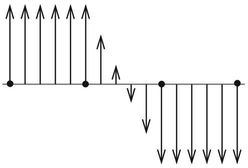


Fig. 7.2 The unit-length constraint is only imposed at the nodes of the triangulation; the linearly interpolated vector field may violate the constraint between two nodes

Proof We have $\mathscr{I}_h|u_h|^2 = 1$ for every $h > 0$ and hence by nodal interpolation estimates and $D^2 u_h|_T = 0$ for every $T \in \mathscr{T}_h$ that

$$\begin{aligned}\big\||u_h|^2 - 1\big\|_{L^2(T)} &= \big\||u_h|^2 - \mathscr{I}_h|u_h|^2\big\|_{L^2(T)} \le ch_T^2 \|D^2|u_h|^2\|_{L^2(T)} \\ &= ch_T^2 \||\nabla u_h|^2\|_{L^2(T)} = ch_T^2 \|\nabla u_h\|_{L^\infty(T)} \|\nabla u_h\|_{L^2(T)}.\end{aligned}$$

The inverse estimate $\|\nabla u_h\|_{L^\infty(T)} \le ch_T^{-1}\|u_h\|_{L^\infty(T)} = ch_T^{-1}$ and a summation over $T \in \mathscr{T}_h$ imply

$$\big\||u_h|^2 - 1\big\| \le ch\|\nabla u_h\|$$

and prove that $|u_h| \to 1$ in $L^2(\Omega)$ as $h \to 0$. Since also $|u_{h'}| \to |u|$ as $h' \to 0$ almost everywhere in Ω for an appropriate subsequence $h' > 0$, we deduce $|u| = 1$ in Ω. □

The proposition motivates minimizing the Dirichlet energy restricted to finite element functions that satisfy the boundary conditions and the unit-length constraint at the nodes of the underlying triangulation.

Theorem 7.5 (Discrete harmonic maps) *Assume that* $\widetilde{u}_{\mathrm{D},h} \in \mathscr{S}^1(\mathscr{T}_h)^m$ *satisfies* $|\widetilde{u}_{\mathrm{D},h}(z)| = 1$ *for all* $z \in \mathscr{N}_h$ *and* $u_{\mathrm{D},h} = \widetilde{u}_{\mathrm{D},h}|_{\Gamma_\mathrm{D}}$. *There exists a minimizer* $u_h \in \mathscr{A}_h$ *for* I *in the set of discrete admissible vector fields*

$$\mathscr{A}_h = \{v_h \in \mathscr{S}^1(\mathscr{T}_h)^m : |v_h(z)| = 1 \textit{ for all } z \in \mathscr{N}_h,\ v_h|_{\Gamma_\mathrm{D}} = u_{\mathrm{D},h}\}.$$

The function $u_h \in \mathscr{A}_h$ *is stationary for* I *in the set of functions in* $\mathscr{A}_h$ *if and only if*

$$(\nabla u_h, \nabla w_h) = 0$$

for all $w_h \in \mathscr{F}_h[u_h]$ *with*

$$\mathscr{F}_h[u_h] = \big\{w_h \in \mathscr{S}^1_\mathrm{D}(\mathscr{T}_h)^m : w_h(z) \cdot u_h(z) = 0 \textit{ for all } z \in \mathscr{N}_h\big\}.$$

Proof The functional I is coercive and continuous on $\mathscr{A}_h$, and this implies the existence of a minimizer. To verify the second statement, let $u_h \in \mathscr{A}_h$ be stationary for I and let $w_h \in \mathscr{F}_h[u_h]$. For every $r \in \mathbb{R}$, we have that $|u_h(z) + rw_h(z)|^2 = |u_h(z)|^2 + r^2|w_h(z)|^2 \ge 1$ for all $z \in \mathscr{N}_h$ and we may define

$$u_h^r = \mathscr{I}_h\Big(\frac{u_h + rw_h}{|u_h + rw_h|}\Big) = \sum_{z\in\mathscr{N}_h} \frac{u_h(z) + rw_h(z)}{|u_h(z) + rw_h(z)|}\varphi_z.$$

For every $z \in \mathscr{N}_h$ a Taylor expansion at $r = 0$ shows that

$$u_h^r(z) = u_h(z) + rw_h(z) + r^2\xi_h(z)$$

for a function $\xi_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m$. Therefore, if u_h is stationary for I, we have

$$0 = \lim_{r\to 0} \frac{1}{r}\big(I(u_h^r) - I(u_h)\big) = (\nabla u_h, \nabla w_h).$$

Conversely, assume that $(\nabla u_h, \nabla w_h) = 0$ for all $w_h \in \mathscr{F}_h[u_h]$. If $(u_h^r)_{r\in(-\varepsilon,\varepsilon)}$ is a continuously differentiable path in $\mathscr{A}_h$ with $w_h^0 = u_h$, then we have

$$u_h^r = u_h + rw_h + \phi(r)\xi_h$$

with a vector field $\xi_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m$, a function ϕ such that $\phi(r)/r \to 0$ as $r \to 0$, and $w_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m$ defined by

$$w_h(z) = \frac{d}{dr}\Big|_{r=0} w_h^r(z).$$

Since $|u_h^r(z)|^2 = 1$ for every $z \in \mathscr{N}_h$ and $r \in (-\varepsilon, \varepsilon)$, we have $w_h(z) \cdot u_h(z) = 0$ for all $z \in \mathscr{N}_h$, i.e., $w_h \in \mathscr{F}_h[u_h]$. This implies

$$I(u_h^r) = I(u_h) + r(\nabla u_h, \nabla w_h) + \phi(r)(\nabla u_h, \nabla \xi_h) + I(rw_h + \phi(r)\xi_h)$$

and thus, using $(\nabla u_h, \nabla w_h) = 0$, we have $\big(I(u_h^r) - I(u_h)\big)/r \to 0$ as $r \to 0$, i.e., $r \mapsto I(u_h^r)$ is stationary at $r = 0$. □

The theorem motivates the following definition.

Definition 7.2 A function $u_h \in \mathscr{A}_h$ is called a *discrete harmonic map* if

$$(\nabla u_h, \nabla w_h) = 0$$

for all $w_h \in \mathscr{F}_h[u_h]$.

Remark 7.8 The space of admissible test functions $\mathscr{F}_h[u_h]$ may be regarded as the tangent space of $\mathscr{A}_h$ at u_h. In particular, a discrete harmonic map is stable with respect to discrete tangential perturbations.

The compactness result of Theorem 7.3 implies the convergence of discrete harmonic maps as $h \to 0$. For ease of presentation we again restrict to the case $m = 3$. The perturbation functionals $\mathscr{R}_h$ in the following theorem model an inexact solution of the discrete problems.

Theorem 7.6 (Discrete compactness) *Let $(u_h)_{h>0} \subset H^1(\Omega;\mathbb{R}^3)$ be a bounded sequence of almost discrete harmonic maps associated to the sequence $(\mathcal{T}_h)_{h>0}$, i.e., for every $h>0$, we have $u_h \in \mathcal{A}_h$ and there exists $\mathcal{R}_h \in H^1_{\mathrm{D}}(\Omega;\mathbb{R}^3)'$ with*

$$(\nabla u_h, \nabla w_h) = \mathcal{R}_h(w_h)$$

for all $w_h \in \mathcal{F}_h[u_h]$. If $\mathcal{R}_h \to 0$ in $H^1_{\mathrm{D}}(\Omega;\mathbb{R}^m)'$ and $u_{\mathrm{D},h} \to u_{\mathrm{D}}$ in $L^2(\Gamma_{\mathrm{D}})$ as $h \to 0$, then every weak accumulation point of $(u_h)_{h>0}$ is a harmonic map.

Proof Let $u \in H^1(\Omega;\mathbb{R}^3)$ be a weak accumulation point of the sequence $(u_h)_{h>0}$ and without loss of generality, assume that the entire sequence converges weakly to u, i.e., $u_h \rightharpoonup u$ in $H^1(\Omega;\mathbb{R}^3)$ as $h \to 0$. Proposition 7.1 shows that $|u| = 1$ almost everywhere in Ω. Moreover, the weak continuity of the trace operator implies that $u|_{\Gamma_{\mathrm{D}}} = u_{\mathrm{D}}$. Given $\phi \in C^\infty(\overline{\Omega};\mathbb{R}^3) \cap H^1_{\mathrm{D}}(\Omega;\mathbb{R}^3)$, set $w_h = \mathcal{I}_h(u_h \times \phi)$. Then $w_h \in \mathcal{S}^1_{\mathrm{D}}(\mathcal{T}_h)^3$ with $w_h(z) \cdot u_h(z) = 0$ for all $z \in \mathcal{N}_h$. An element-wise nodal interpolation estimate and $D^2 u_h|_T = 0$ for every $T \in \mathcal{T}_h$ show that

$$\begin{aligned}\|\nabla(w_h - u_h \times \phi)\|_{L^2(T)} &\le c h_T \|D^2(u_h \times \phi)\|_{L^2(T)} \\ &\le c h_T \big(\|\nabla u_h\|_{L^2(T)} \|\nabla \phi\|_{L^\infty(T)} + \|u_h\|_{L^\infty(T)} \|\nabla \phi\|_{L^2(T)}\big).\end{aligned}$$

This implies that $\|\nabla w_h\| \le c$ and $w_h - u_h \times w \to 0$ in $H^1(\Omega;\mathbb{R}^3)$ as $h \to 0$. Therefore, we have

$$\mathcal{R}_h(w_h) = (\nabla u_h, \nabla w_h) = (\nabla u_h, \nabla[u_h \times \phi]) + (\nabla u_h, \nabla[w_h - u_h \times \phi])$$

with

$$(\nabla u_h, \nabla[w_h - u_h \times \phi]) \to 0$$

as $h \to 0$. For the other term on the right-hand side, we have

$$\sum_{\ell=1}^d (\partial_\ell u_h, \partial_\ell[u_h \times \phi]) = \sum_{\ell=1}^d (\partial_\ell u_h, u_h \times \partial_\ell \phi)$$

and since $u_h \to u$ in $L^2(\Omega;\mathbb{R}^3)$ and $\nabla u_h \rightharpoonup \nabla u$ in $L^2(\Omega;\mathbb{R}^{3\times 3})$ as $h \to 0$, we deduce that

$$0 = \lim_{h\to 0} (\nabla u_h, \nabla[u_h \times \phi]) = \sum_{\ell=1}^d (\partial_\ell u, u \times \partial_\ell \phi).$$

This proves that u is a harmonic map. □

7.2.2 Iterative Computation

The iterative computation of discrete harmonic maps is based on the computation of tangential corrections that define a new approximation after a node-wise projection onto the unit sphere. The following algorithm may be regarded as a discrete version of the H^1-flow for harmonic maps which is formally defined as

$$(\nabla \partial_t u, \nabla w) = -(\nabla u, \nabla w) + (|\nabla u|^2 u, w).$$

For w with $w \cdot u = 0$, the second term on the right-hand side disappears. Moreover, we have $\partial_t u \cdot u = 0$ if $|u(t,x)| = 1$ for almost every $(t,x) \in [0,T] \times \Omega$. We employ a semi-implicit discretization of this problem to compute approximations v_h^k of $\partial_t u(t_k)$ to find discrete harmonic maps with bounded energy. In particular, the linearized constraint will be treated explicitly, which leads to linear systems of equations in every time-step. The approach is illustrated in Fig. 7.3.

Algorithm 7.1 (*Discrete H^1-flow* [1]) Let $u_h^0 \in \mathscr{A}_h$, $\theta \in [0,1]$, and $\tau > 0$ and define the sequence $(u_h^k)_{k=0,1,\ldots} \subset \mathscr{A}_h$ by computing $v_h^k \in \mathscr{F}_h[u_h^{k-1}]$ such that

$$(\nabla v_h^k, \nabla w_h) + (\nabla[u_h^{k-1} + \theta\tau v_h^k], \nabla w_h) = 0$$

for all $w_h \in \mathscr{F}_h[u_h^{k-1}]$ and setting

$$u_h^k = \sum_{z \in \mathscr{N}_h} \frac{u_h^{k-1}(z) + \tau v_h^k(z)}{|u_h^{k-1}(z) + \tau v_h^k(z)|} \varphi_z$$

until $\|\nabla v_h^k\| \le \varepsilon_{\text{stop}}$.

Proposition 7.2 (Termination I) *Assume that $\mathscr{T}_h$ is weakly acute. The iterates $(u_h^k)_{k=0,1,\ldots} \subset \mathscr{A}_h$ of Algorithm 7.1 are well defined and satisfy*

$$\frac{1}{2}\|\nabla u_h^L\|^2 + (2 + 2\tau\theta - \tau)\frac{\tau}{2}\sum_{k=1}^{L} \|\nabla v_h^k\|^2 \le \frac{1}{2}\|\nabla u_h^0\|^2$$

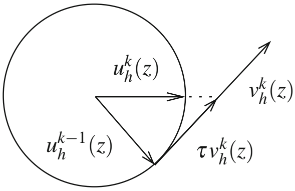


Fig. 7.3 The iteration of Algorithm 7.1 computes corrections v_h^k in the tangent space of the unit sphere at the current iterate u_h^{k-1} and then employs a projection onto the unit sphere to define the update u_h^k

for every $L \geq 1$. *In particular, if* $\tau(1-2\theta) \leq 2$, *then the iteration terminates and the output* $u_h^* \in \mathscr{A}_h$ *satisfies*

$$(\nabla u_h^*, \nabla w_h) = \mathscr{R}_h(w_h)$$

for all $w_h \in \mathscr{F}_h[u_h^*]$ *and* $\|\mathscr{R}_h\|_{H^1_{\mathrm{D}}(\Omega;\mathbb{R}^m)'} \leq (1+\theta\tau)\varepsilon_{\mathrm{stop}}$.

Proof Given $u_h^{k-1} \in \mathscr{A}_h$, the space $\mathscr{F}_h[u_h^{k-1}]$ is a closed subspace of $\mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m$ and the Lax–Milgram lemma implies the existence of a uniquely defined $v_h^k \in \mathscr{F}_h[u_h^{k-1}]$ with

$$(\nabla v_h^k, \nabla w_h) + (\nabla[u_h^{k-1} + \theta\tau v_h^k], \nabla w_h) = 0$$

for all $w_h \in \mathscr{F}_h[u_h^{k-1}]$. Since $|u_h^{k-1}(z)| = 1$ and $v_h^k(z) \cdot u_h^{k-1}(z) = 0$ for all $z \in \mathscr{N}_h$, we have $|u_h^{k-1}(z) + \tau v_h^k(z)| \geq 1$ and $u_h^k \in \mathscr{A}_h$ is well defined. The mapping

$$F : s \mapsto \begin{cases} s/|s| & \text{if } |s| \geq 1, \\ s & \text{if } |s| \leq 1 \end{cases}$$

is Lipschitz continuous with $\|DF\|_{L^\infty(\mathbb{R}^m)} = 1$ so that Proposition 3.2 implies

$$\|\nabla u_h^k\| \leq \|\nabla(u_h^{k-1} + \tau v_h^k)\|.$$

The choice of $w_h = v_h^k$ in the equation of Algorithm 7.1 and the formula $2\tau(a + \theta\tau b)b = (a+\tau b)^2 - a^2 + \tau^2(2\theta-1)b^2$ show that

$$\|\nabla v_h^k\|^2 + \frac{1}{2\tau}\|\nabla(u_h^{k-1} + \tau v_h^k)\|^2 - \frac{1}{2\tau}\|\nabla u_h^{k-1}\|^2 + \frac{\tau}{2}(2\theta-1)\|\nabla v_h^k\|^2 = 0.$$

A combination with the bound for $\|\nabla u_h^k\|$ and a multiplication by τ, together with a summation over $k = 1, 2, \dots, L$, imply

$$\frac{1}{2}\|\nabla u_h^L\|^2 + (2 + 2\tau\theta - \tau)\frac{\tau}{2}\sum_{k=1}^{L}\|\nabla v_h^k\|^2 \leq \frac{1}{2}\|\nabla u_h^0\|^2.$$

This yields that $\|\nabla v_h^K\| \leq \varepsilon$ for $K \geq 0$ sufficiently large and the functions $u_h^* = u_h^{K-1}$ and $v_h^* = v_h^K$ satisfy

$$(\nabla u_h^*, \nabla w_h) = -(1+\theta\tau)(\nabla v_h^*, \nabla w_h)$$

for all $w_h \in \mathscr{F}_h[u_h^*]$. Setting $\mathscr{R}_h(w) = -(1+\theta\tau)(\nabla v_h^*, \nabla w)$ for $w \in H^1_{\mathrm{D}}(\Omega;\mathbb{R}^m)$ proves the assertion. □

Remarks 7.9 (i) The proof of the proposition shows that we have the local energy decay property $\|\nabla u_h^k\| \leq \|\nabla u_h^{k-1}\|$ for all $k \geq 1$.

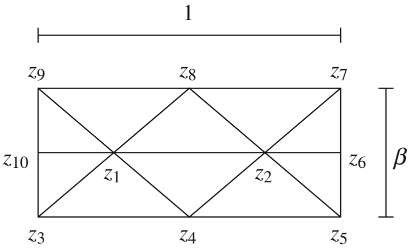


Fig. 7.4 A triangulation $\mathscr{T}_h$ that is weakly acute if and only if $\beta \geq 1/2$

(ii) Note that for all choices of θ the large step size $\tau = 1$ leads to a stable and convergent iterative scheme.

The acuteness property is necessary in general to guarantee that the projection step is stable in the sense that $\|\nabla u_h^k\| \leq \|\nabla[u_h^{k-1} + \tau v_h^k]\|$.

Proposition 7.3 (Necessity of acuteness) *For $\beta > 0$, let $\mathscr{T}_h$ be the triangulation of $\Omega = (0,1) \times (0,\beta)$ shown in Fig. 7.4, and let $\tau > 0$. Let $u_h \in \mathscr{S}^1(\mathscr{T}_h)^m$ and $v_h \in \mathscr{F}_h[u_h]$, be defined by $u_h(z_j) = e_1$ and $v_h(z_j) = 0$ for $j = 3, 4, \ldots, 10$, and*

$$\begin{aligned} u_h(z_1) &= e_1, & u_h(z_2) &= -e_1, \\ v_h(z_1) &= -(s/\tau)e_2, & v_h(z_2) &= 0, \end{aligned}$$

where $s = 1/2 - \beta$ and e_ℓ denotes the ℓ-th canonical basis vector in $\mathbb{R}^m$. Then for $P\widetilde{u}_h \in \mathscr{S}^1(\mathscr{T}_h)^m$ defined with $\widetilde{u}_h = u_h + \tau v_h$ by

$$P\widetilde{u}_h(z) = \frac{\widetilde{u}_h(z)}{|\widetilde{u}_h(z)|}$$

for all $z \in \mathscr{N}_h$, we have $\|\nabla P\widetilde{u}_h\| \leq \|\nabla \widetilde{u}_h\|$ if and only if $\mathscr{T}_h$ is weakly acute, i.e., if and only if $\beta \geq 1/2$.

Proof Since $|\widetilde{u}_h(z)| \geq 1$ for all $z \in \mathscr{N}_h$, Proposition 3.2 implies that $\|\nabla P\widetilde{u}_h\| \leq \|\nabla \widetilde{u}_h\|$ if $\mathscr{T}_h$ is weakly acute and this is the case if and only if $\beta \geq 1/2$. Suppose that $\beta < 1/2$. Then with the entries A_{jk}, $j, k = 1, 2, \ldots, 10$, of the stiffness matrix and the identity $\widetilde{u}_h(z_j) = P\widetilde{u}_h(z_j)$ for $j = 2, 3, 4, \ldots, 10$, the representation of $\|\nabla w_h\|^2$ in terms of the nodal values of w_h and the entries of A, cf. the proof of Proposition 3.2, we have that

$$\begin{aligned} \delta^2 = \|\nabla \widetilde{u}_h\|^2 - \|\nabla P\widetilde{u}_h\|^2 &= -\frac{1}{2} \sum_{j,k=1}^{10} A_{jk} \big(|\widetilde{u}_h(z_j) - \widetilde{u}_h(z_k)|^2 \\ &\qquad - |P\widetilde{u}_h(z_j) - P\widetilde{u}_h(z_k)|^2 \big) \\ &= -\sum_{j=2}^{10} A_{1j} \big(|\widetilde{u}_h(z_j) - \widetilde{u}_h(z_1)|^2 \end{aligned}$$

$$- |P\widetilde{u}_h(z_j) - P\widetilde{u}_h(z_1)|^2\big).$$

We have $|\widetilde{u}_h(z_1) - \widetilde{u}_h(z_2)|^2 = 4 + s^2$ and $|\widetilde{u}_h(z_j) - \widetilde{u}_h(z_1)|^2 = s^2$ and

$$\begin{aligned} t_1^2 &= |P\widetilde{u}_h(z_1) - P\widetilde{u}_h(z_2)|^2 = 2 + 2/(1+s^2)^{1/2}, \\ t_2^2 &= |P\widetilde{u}_h(z_j) - P\widetilde{u}_h(z_2)|^2 = 2 - 2/(1+s^2)^{1/2} \end{aligned}$$

for $j = 3, 4, \ldots, 10$. Since $\sum_{j=1}^{10} A_{1j} = 0$ we have $\sum_{j=3}^{10} A_{1j} = -A_{11} - A_{22}$ and hence

$$\begin{aligned} \delta^2 &= (s^2 - t_2^2)(A_{11} + A_{12}) - A_{12}(4 + s^2 - t_1^2) \\ &= A_{11}(s^2 - t_2^2) - A_{12}(4 + t_2^2 - t_1^2). \end{aligned}$$

Direct calculations show that

$$A_{11} = (12\beta^2 + 5)/(4\beta), \quad A_{12} = (1 - 4\beta^2)/(4\beta).$$

With $\phi(s) = (1+s^2)^{1/2} - 1 - s^2/2$ and $\beta^2 = 1/4 - s + s^2$ we verify that

$$\begin{aligned} 4\beta(1+s^2)^{1/2}\delta^2 &= \big(12\beta^2 + 5\big)\big(s^4/2 + s^2\phi(s) - 2\phi(s)\big) - \big(1 - 4\beta^2\big)\big(2s^2 + 4\phi(s)\big) \\ &= \big(8 - 12s + 12s^2\big)\big(s^4/2 + s^2\phi(s) - 2\phi(s)\big) \\ &\quad - 16(s - s^2)\big(s^2/2 + \phi(s)\big) \\ &= -8s^3 + 12s^4 - 6s^5 + 6s^6 + \phi(s)\big(-16s - 12s^3 + 12s^4\big) \\ &= -6s^3(1-2s) - 6s^5(1-s) + 4s\phi(s)\big(2 - 3s^2 + 3s^3\big) \\ &\quad - 2\big(s^3 + 8\phi(s)\big). \end{aligned}$$

Since $0 < s < 1/2$ and $\phi(s) < 0$, the first three terms on the right-hand side are negative. The estimate $-s^4/8 \le \phi(s)$ implies that the last term on the right-hand side is nonpositive. This shows $\delta < 0$ if $\beta < 1/2$ and proves the assertion. □

7.2.3 Projection-Free Iteration

The acuteness condition of Proposition 7.2 is restrictive if $d = 3$ but allows for large step sizes. In the continuous situation we have that the identity $u \cdot \partial_t u = 0$ implies that the initial length is preserved. In the discrete setting a semi-implicit discretization of this orthogonality leads to approximations that violate the constraint when the projection step is omitted, cf. Fig. 7.5.

Algorithm 7.2 (*H^1-flow without projection*) Let $u_h^0 \in \mathscr{A}_h$, $\tau > 0$, and define the sequence $(u_h^k)_{k=0,1,\ldots} \subset \mathscr{S}^1(\mathscr{T}_h)^m$ by computing $v_h^k \in \mathscr{F}_h[u_h^{k-1}]$ such that

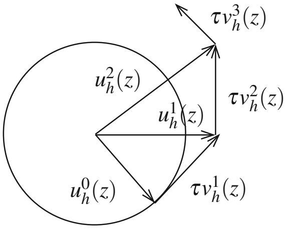


Fig. 7.5 Omitting the projection step in the semi-implicit H^1-flow leads to approximations that violate the unit-length constraint; the corresponding error in $L^1(\Omega)$ is independent of the number of iterations and controlled by the step size

$$(\nabla v_h^k, \nabla w_h) + (\nabla[u_h^{k-1} + \tau v_h^k], \nabla w_h) = 0$$

for all $w_h \in \mathscr{F}_h[u_h^{k-1}]$ and setting

$$u_h^k = u_h^{k-1} + \tau v_h^k$$

until $\|\nabla v_h^k\| \le \varepsilon_{\text{stop}}$.

The following proposition shows that the violation of the constraint is independent of the number of iterations and controlled by the step size.

Proposition 7.4 (Termination II) *The iterates* $(u_h^k)_{k=0,1,\dots} \subset \mathscr{S}^1(\mathscr{T}_h)^m$ *of Algorithm*7.2 *satisfy* $u_h^k|_{\Gamma_\mathrm{D}} = u_{\mathrm{D},h}$ *for* $k = 0, 1, \dots$ *and*

$$\frac{1}{2}\|\nabla u_h^L\|^2 + (2+\tau)\frac{\tau}{2}\sum_{k=1}^{L}\|\nabla v_h^k\|^2 = \frac{1}{2}\|\nabla u_h^0\|^2$$

for every $L \ge 1$. *Moreover, we have every* $L \ge 1$ *that*

$$\big\|\mathscr{I}_h\big[|u_h^L|^2\big] - 1\big\|_{L^1(\Omega)} \le c\tau\|\nabla u_h^0\|^2.$$

Proof Due to the Lax–Milgram lemma the iteration is well-defined and the choice $w_h^k = v_h^{k+1}$ shows, using the formula $2\tau(a+\tau b)b = (a+\tau b)^2 - a^2 + \tau^2 b^2$, that

$$\frac{2+\tau}{2}\|\nabla v_h^k\|^2 + \frac{1}{2\tau}\|\nabla u_h^k\|^2 - \frac{1}{2\tau}\|\nabla u_h^{k-1}\|^2 = 0$$

which implies the first asserted estimate. For every $z \in \mathscr{N}_h$, we have

$$|u_h^k(z)|^2 - 1 = |u_h^{k-1}(z)|^2 + \tau^2|v_h^k(z)|^2 - 1$$

and inductively with $|u_h^0(z)| = 1$, we find that

$$|u_h^L(z)|^2 - 1 = \tau^2\sum_{k=1}^{L}|v_h^k(z)|^2.$$

The discrete norm equivalences of Lemma 3.4 yield

$$(1/c)\big\|\mathscr{I}_h\big[|u_h^L|^2\big]-1\big\|_{L^1(\Omega)} \le \sum_{z\in\mathcal{N}_h} h_z^d\big||u_h^L(z)|^2-1\big|$$
$$\le \tau^2\sum_{k=1}^L\sum_{z\in\mathcal{N}_h} h_z^d|v_h^k(z)|^2 \le c\tau^2\sum_{k=1}^L\|v_h^k\|^2.$$

Poincaré's inequality and the first estimate of the proposition imply

$$\big\|\mathscr{I}_h\big[|u_h^L|^2\big]-1\big\|_{L^1(\Omega)} \le c\tau^2\sum_{k=1}^L\|\nabla v_h^k\|^2 \le c\tau\|\nabla u_h^0\|^2,$$

which proves the proposition. □

We conclude the discussion with a lemma which shows that the approximate treatment of the constraint at the nodes implies that it is satisfied by accumulation points in the limit $(h,\tau)\to 0$.

Lemma 7.2 (Constraint approximation) *If $(u_h)_{h>0}$ is a bounded sequence in $H^1(\Omega;\mathbb{R}^m)$ such that $u_h\in\mathscr{S}^1(\mathscr{T}_h)^m$ for all $h>0$, $u_h\to u$ in $L^2(\Omega;\mathbb{R}^m)$ for some $u\in H^1(\Omega;\mathbb{R}^m)$ as $h\to 0$, and*

$$\|\mathscr{I}_h\big[|u_h|^2\big]-1\|_{L^1(\Omega)}\to 0$$

as $h\to 0$, then we have $|u|^2=1$ almost everywhere in Ω.

Proof Two applications of the triangle inequality show that

$$\||u|^2-1\|_{L^1(\Omega)}$$
$$\le \||u|^2-|u_h|^2\|_{L^1(\Omega)}+\||u_h|^2-\mathscr{I}_h\big[|u_h|^2\big]\|_{L^1(\Omega)}+\|\mathscr{I}_h\big[|u_h|^2\big]-1\|_{L^1(\Omega)}.$$

Due to the assumptions of the lemma we have that the third term on the right-hand side tends to zero as $h\to 0$. Since

$$\||u|^2-|u_h|^2\|_{L^1(\Omega)}\le\|u-u_h\|\|u+u_h\|$$

we have that also the first term on the right-hand side vanishes as $h\to 0$. We use Hölder's inequality and a nodal interpolation estimate to verify that for every $T\in\mathscr{T}_h$, we have

$$\||u_h|^2-\mathscr{I}_h\big[|u_h|^2\big]\|_{L^1(T)} \le ch_T^{d/2}\||u_h|^2-\mathscr{I}_h\big[|u_h|\big]\|_{L^2(T)}$$
$$\le ch_T^{d/2}h_T^2\|D^2|u_h|^2\|_{L^2(T)} \le ch_T^2\|\nabla u_h\|_{L^2(T)}^2.$$

With a summation over $T \in \mathscr{T}_h$ we deduce for the second term that

$$\| |u_h|^2 - \mathscr{I}_h\big[|u_h|^2\big] \|_{L^1(\Omega)} \leq ch^2 \|\nabla u_h\|^2.$$

Since the upper bound vanishes as $h \to 0$, this implies that $|u|^2 = 1$. □

7.2.4 *Other Target Manifolds*

The ideas outlined above can be generalized to approximate harmonic maps into target manifolds other than the unit sphere. We let $\mathscr{M} \subset \mathbb{R}^m$ be an $(m-1)$-dimensional C^2-submanifold and let $T_p\mathscr{M}$ denote the tangent space at $p \in \mathscr{M}$. Moreover, we let $\pi_{\mathscr{M}} : U_\delta(\mathscr{M}) \to \mathscr{M}$ be the nearest neighbor projection onto $\mathscr{M}$ which is uniquely defined in a neighborhood $U_\delta(\mathscr{M}) = \{q \in \mathbb{R}^m : \operatorname{dist}(p, \mathscr{M}) < \delta\}$ of $\mathscr{M}$ for some $\delta > 0$. The function $\pi_{\mathscr{M}}$ satisfies $|\pi_{\mathscr{M}}(q) - q| = \inf_{p\in\mathscr{M}} |p-q|$ for all $q \in U_\delta(\mathscr{M})$. If $\mathscr{M} = \partial\mathscr{C}$ for a convex set $\mathscr{C} \subset \mathbb{R}^m$, then $\pi_{\mathscr{M}}$ is well defined in $\mathbb{R}^m \setminus \mathscr{C}$.

Definition 7.3 Given $\widetilde{u}_{\mathrm{D},h} \in \mathscr{S}^1(\mathscr{T}_h)^m$ with $\widetilde{u}_{\mathrm{D},h}(z) \in \mathscr{M}$ for all $z \in \mathscr{N}_h$ set

$$\mathscr{A}_h = \{u_h \in \mathscr{S}^1(\mathscr{T}_h)^m : u_h|_{\Gamma_\mathrm{D}} = \widetilde{u}_{\mathrm{D},h}|_{\Gamma_\mathrm{D}} \text{ and } u_h(z) \in \mathscr{M} \text{ for all } z \in \mathscr{N}_h\big\}$$

and for $u_h \in \mathscr{A}_h$, let

$$\mathscr{F}_h[u_h] = \big\{v_h \in \mathscr{S}^1_\mathrm{D}(\mathscr{T}_h)^m : v_h(z) \in T_{u_h(z)}\mathscr{M} \text{ for all } z \in \mathscr{N}_h\big\}.$$

With these definitions we can define the following generalization of Algorithm 7.1.

Algorithm 7.3 (*H^1-flow for general target manifolds*) Let $u_h^0 \in \mathscr{A}_h$ and $\tau > 0$ and define the sequence $(u_h^k)_{k=0,1,\dots} \in \mathscr{A}_h$ by computing $v_h^k \in \mathscr{F}_h[u_h^{k-1}]$ such that

$$(\nabla v_h^k, \nabla w_h) + (\nabla[u_h^{k-1} + \tau v_h^k], \nabla w_h) = 0$$

for all $w_h \in \mathscr{F}_h[u_h^{k-1}]$ and setting

$$u_h^k = \sum_{z\in\mathscr{N}_h} \pi_{\mathscr{M}}\big(u_h^{k-1}(z) + \tau v_h^k(z)\big)\varphi_z$$

until $\|\nabla v_h^k\| \leq \varepsilon_{\mathrm{stop}}$.

Remarks 7.10 (i) Well-posedness of the algorithm requires that τ be sufficiently small so that $u_h^{k-1}(z) + \tau v_h^k(z) \in U_\delta(\mathscr{M})$ for all $z \in \mathscr{N}_h$, cf. Fig. 7.6. If $\mathscr{M} = \partial\mathscr{C}$ for a convex set $\mathscr{C}$, then this is always satisfied.
(ii) A stability proof employs an expansion of $\pi_{\mathscr{M}}$ and the fact that $D\pi_{\mathscr{M}}(s)|_{T_s\mathscr{M}} = \mathrm{id}_{T_s\mathscr{M}}$ provided that $\mathscr{M}$ is a C^3-submanifold.

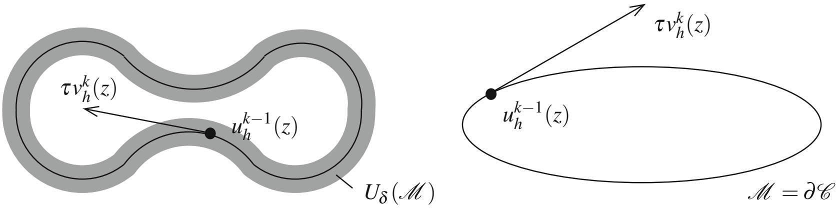


Fig. 7.6 The projection of $u_h^{k-1}(z) + \tau v_h^k(z)$ onto the target manifold is in general only well defined within a tubular neighborhood of $\mathscr{M}$ in the case of nonconvex manifolds and a step-size restriction needs to be imposed (*left*); for boundaries of convex sets no restriction on the step size is required (*right*)

(iii) The projection step can be omitted if an appropriate version of a shifted tangent space is available, e.g., if $\mathscr{M} = g^{-1}(\{0\})$ for an appropriate function $g : \mathbb{R}^m \to \mathbb{R}$.

7.2.5 Practical Realization

The implementation of Algorithm 7.1 requires working with discrete vector fields $u_h \in \mathscr{S}^1(\mathscr{T}_h)^m$ which are given by

$$u_h = \sum_{z \in \mathscr{N}_h} u_z \varphi_z$$

with coefficients $u_z = u_h(z) \in \mathbb{R}^m$ for all $z \in \mathscr{N}_h$. The function u_h will be identified with the vector $U \in \mathbb{R}^{mL}$ defined by

$$U = \begin{bmatrix} u_{z_1} \\ u_{z_2} \\ \vdots \\ u_{z_L} \end{bmatrix} \in \mathbb{R}^{mL}$$

with $L = \#\mathscr{N}_h$. The constraint $u_h(z) \cdot v_h(z) = 0$ for all $z \in \mathscr{N}_h$ for a vector field $v_h \in \mathscr{S}^1(\mathscr{T}_h)^m$ is then equivalently imposed by $B_U V = 0$ with the matrix $B_U \in \mathbb{R}^{L \times L}$ defined through

$$B_U = \begin{bmatrix} u_{z_1}^\top & 0 & & \\ 0 & u_{z_2}^\top & 0 & \\ & 0 & \ddots & 0 \\ & & 0 & u_{z_L}^\top \end{bmatrix}$$

so that $B_U V = [u_{z_1} \cdot v_{z_1}, u_{z_2} \cdot v_{z_2}, \dots, u_{z_L} \cdot v_{z_L}]^\top$. The solution of the linearly constrained linear problems is based on the fact that we have

$$B_U V = 0, \quad W^\top S_m V = W^\top b \ \text{ for all } W \in \ker B$$

if and only if there exists $\Lambda \in \mathbb{R}^L$ such that

$$\begin{bmatrix} S_m & B_U^\top \\ B_U & 0 \end{bmatrix} \begin{bmatrix} V \\ \Lambda \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix},$$

where S_m is the $P1$ finite element stiffness matrix for vector fields with m components. A MATLAB implementation is shown in Fig. 7.7.

7.3 Approximation of Constrained Evolution Problems

The iterative schemes discussed above are discrete H^1-gradient flows for harmonic maps and can be modified to provide approximations of the L^2-gradient flow of harmonic maps. We show that this leads to convergent approximations of the harmonic map heat flow. In addition to this we analyze discretizations that preserve the constraint without an explicit correction of the iterates. We also discuss the application of the developed techniques to a hyperbolic problem. The presentation is based on results from [2, 4, 7].

7.3.1 Harmonic Map Heat Flow

The harmonic map heat flow is the L^2-gradient flow for the Dirichlet energy that is constrained to unit-length vector fields. In the strong form it seeks a function $u : [0, T] \times \Omega \to \mathbb{R}^m$ such that $|u| = 1$ in $[0, T] \times \Omega$ and

$$\partial_t u - \Delta u = |\nabla u|^2 u, \quad u|_{\Gamma_\mathrm{D}} = u_\mathrm{D}, \quad \partial_n u|_{\Gamma_\mathrm{N}} = 0, \quad u(0) = u_0,$$

where Γ_D may be empty. The following proposition provides useful equivalent characterizations for the practically relevant case $m = 3$.

Proposition 7.5 (Equivalent formulations) *The following formulations are equivalent for a function* $u \in H^1([0, T]; L^2(\Omega; \mathbb{R}^3)) \cap L^\infty([0, T]; H^1(\Omega; \mathbb{R}^3))$ *satisfying* $|u(t, x)| = 1$ *for almost every* $(t, x) \in [0, T] \times \Omega$*:*
(i) *For almost every* $t \in [0, T]$ *and every* $w \in H^1_\mathrm{D}(\Omega; \mathbb{R}^3) \cap L^\infty(\Omega; \mathbb{R}^3)$*, we have*

$$(\partial_t u, w) + (\nabla u, \nabla w) = (|\nabla u|^2 u, w).$$

```
function h1_flow_hm(d,red)
[c4n,n4e,Db,Nb] = triang_cube(d); Db = [Db;Nb]; c4n = c4n-.5;
for j = 1:red
    [c4n,n4e,Db,Nb,¬,¬] = red_refine(c4n,n4e,Db,Nb);
end
theta = 1/2; tau = 2; eps_stop = 1e-4; nC = size(c4n,1);
dNodes = unique(Db); fNodes = setdiff(1:nC,dNodes);
FNodes = [3*fNodes-2,3*fNodes-1,3*fNodes-0];
nDb = size(dNodes,1); nF = size(fNodes,2);
[s,¬,¬,¬] = fe_matrices(c4n,n4e); SSS = sparse(3*nC,3*nC);
for k = 1:3
    idx = k:3:3*nC; SSS(idx,idx) = s;
end
u = zeros(3*nC,1);
for j = 1:nC
    u(3*j-[2,1,0]) = u_0(c4n(j,:));
end
for j = 1:nDb
    u(3*dNodes(j)-[2,1,0]) = u_D(c4n(dNodes(j),:));
end
Flist = [FNodes,3*nC+fNodes];
norm_corr = 1;
while norm_corr > eps_stop
    B = sparse(nC,3*nC);
    for j = 1:nC
        B(j,3*j-[2,1,0]) = u(3*j-[2,1,0])';
    end
    X = [SSS,B';B,sparse(nC,nC)];
    b = [-(1+theta*tau)^(-1)*SSS*u;zeros(nC,1)];
    x = X(Flist,Flist)\b(Flist);
    v = zeros(3*nC,1);
    v(FNodes) = x(1:3*nF); tu = u+tau*v;
    norm_corr = sqrt(v'*SSS*v);
    for j = 1:nC
        u(3*j-[2,1,0]) = tu(3*j-[2,1,0])/norm(tu(3*j-[2,1,0]));
    end
    % u = tu;
    show_p1_field(c4n,u); axis square; view(30,30); drawnow;
end

function val = u_D(x)
val = [x/norm(x),zeros(1,3-size(x,2))];

function val = u_0(x)
val_tmp = rand(1,3)-.5;
val = val_tmp/norm(val_tmp);

function show_p1_field(c4n,u)
[nC,d] = size(c4n); X = [c4n,zeros(nC,3-d)];
quiver3(X(:,1),X(:,2),X(:,3),u(1:3:3*nC),u(2:3:3*nC),u(3:3:3*nC));
```

Fig. 7.7 Iterative approximation of harmonic maps into the sphere S^2 incorporating a projection step which can be deactivated by uncommenting the command `u = tu;`

(ii) *For almost every* $t \in [0, T]$ *and every* $w \in H^1_{\mathrm{D}}(\Omega; \mathbb{R}^3)$ *with* $w(x) \cdot u(t, x) = 0$ *for almost every* $x \in \Omega$*, we have*

$$(\partial_t u, w) + (\nabla u, \nabla w) = 0.$$

(iii) *For almost every* $t \in [0, T]$ *and every* $\phi \in H^1_{\mathrm{D}}(\Omega; \mathbb{R}^3) \cap L^\infty(\Omega; \mathbb{R}^3)$*, we have*

$$(\partial_t u, \phi) - (\nabla u, \nabla[u \times (u \times \phi)]) = 0.$$

Proof The proof is similar to the proof of Lemma 7.1. Assume that formulation (i) is satisfied. If $w(x) \cdot u(x, t) = 0$, then the right-hand side vanishes and a truncation argument shows that formulation (ii) holds. Using the identity $w = u \times (u \times \phi) = u(u\cdot\phi) - \phi$ implies the equivalence of (i) and (iii). Finally, (iii) follows from choosing $w = u \times (u \times \phi)$ in (ii) and noting that $\partial_t u \cdot u = 0$. □

Remark 7.11 The equivalence of (i) and (ii) can also be established for functions with values in $\mathbb{R}^m$ with $m \neq 3$.

7.3.2 Semi-implicit, Linear Schemes

The L^2-flow of harmonic maps can be approximated by replacing the H^1-inner product in Algorithm 7.1 by the L^2-inner product. As in that algorithm, the projection step can be omitted leading to a violation of the unit length constraint that is controlled by the step size independently of the number of iterations or time steps. As above we denote

$$\mathscr{A}_h = \big\{ v_h \in \mathscr{S}^1(\mathscr{T}_h)^m : |v_h(z)| = 1 \text{ for all } z \in \mathscr{N}_h,\ v_h|_{\Gamma_{\mathrm{D}}} = u_{\mathrm{D},h} \big\}$$

and given any $u_h \in \mathscr{S}^1(\mathscr{T}_h)^m$, we denote

$$\mathscr{F}_h[u_h] = \big\{ v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m : v_h(z) \cdot u_h(z) = 0 \text{ for all } z \in \mathscr{N}_h \big\}.$$

Here and throughout the following the set Γ_{D} may be empty.

Algorithm 7.4 (*Discrete L^2-flow with optional projection*) Let $u_h^0 \in \mathscr{A}_h$, $\theta \in [0,1]$, and $\tau > 0$ and define the sequence $(u_h^k)_{k=0,\dots,K} \subset \mathscr{S}^1(\mathscr{T}_h)^m$ for $K = \lceil T/\tau \rceil$ by computing for $k = 1, 2, \dots, K$ the function $v_h^k \in \mathscr{F}_h[u_h^{k-1}]$ such that

$$(v_h^k, w_h) + (\nabla[u_h^{k-1} + \theta\tau v_h^k], \nabla w_h) = 0$$

for all $w_h \in \mathscr{F}_h[u_h^{k-1}]$ and setting $\widetilde{u}_h^k = u_h^{k-1} + \tau v_h^k$ and

$$u_h^k = \widetilde{u}_h^k \qquad \text{or} \qquad u_h^k = \sum_{z \in \mathscr{N}_h} \frac{\widetilde{u}_h^k(z)}{|\widetilde{u}_h^k(z)|} \varphi_z.$$

We discuss the stability properties of the algorithm for the case $\theta = 1$.

Proposition 7.6 (Stability) *Let $(u_h^k)_{k=0,\dots,K} \subset \mathscr{S}^1(\mathscr{T}_h)^m$ be the iterates of Algorithm 7.4 for $\theta = 1$.*
(i) *If the projection is omitted, then we have $v_h^k = d_t u_h^k$ for $k = 1, 2, \dots, K$ and for $L = 1, 2, \dots, K$*

$$\frac{1}{2}\|\nabla u_h^L\|^2 + \tau \sum_{k=1}^{L} \Big(\frac{\tau}{2}\|\nabla d_t u_h^k\|^2 + \|d_t u_h^k\|^2\Big) = \frac{1}{2}\|\nabla u_h^0\|^2,$$

$$\big\|\mathscr{I}_h\big[|u_h^L|^2\big] - 1\big\|_{L^1(\Omega)} \le c_0 \tau.$$

(ii) *If the projection step is included and if $\mathscr{T}_h$ is weakly acute, then $u_h^k \in \mathscr{A}_h$ for $k = 0, 1, \dots, K$ and for every $L = 1, 2, \dots, K$, we have*

$$\frac{1}{2}\|\nabla u_h^L\|^2 + \tau \sum_{k=1}^{L} \Big(\frac{\tau}{2}\|\nabla v_h^k\|^2 + \|v_h^k\|^2\Big) \le \frac{1}{2}\|\nabla u_h^0\|^2,$$

$$\tau \sum_{k=1}^{L} \|v_h^k - d_t u_h^k\|_{L^1(\Omega)} \le c_0 \tau.$$

Proof The well-posedness of Algorithm 7.4 follows as in the case of Algorithm 7.1 with the help of the Lax–Milgram lemma and the fact that $|\widetilde{u}_h^k(z)| \ge 1$ for all $k \ge 1$ and $z \in \mathscr{N}_h$.
(i) Assume that the projection step in Algorithm 7.4 is omitted. We then have $v_h^k = d_t u_h^k$ and the choice of $w_h = d_t u_h^k$ yields

$$\|d_t u_h^k\|^2 + \frac{d_t}{2}\|\nabla u_h^k\|^2 + \frac{\tau}{2}\|\nabla d_t u_h^k\|^2 = 0.$$

A summation over $k = 1, 2, \dots, L$ and multiplication by τ prove the stability estimate. For all $z \in \mathscr{N}_h$ and $k = 1, 2, \dots, L$, we have

$$|u_h^k(z)|^2 = |u_h^{k-1}(z) + \tau d_t u_h^k(z)|^2 = |u_h^{k-1}(z)|^2 + \tau^2 |d_t u_h^k(z)|^2$$

and inductively it follows with $|u_h^0(z)| = 1$ that

$$|u_h^L(z)|^2 - 1 = \tau^2 \sum_{k=1}^{L} |d_t u_h^k(z)|^2.$$

Multiplication by h_z^d the norm equivalences of Lemma 3.4, and the stability estimate imply, as in the proof of Proposition 7.4, that

$$\big\|\mathscr{I}_h\big[|u_h^L|^2\big]-1\big\|_{L^1(\Omega)} \le c\tau^2\sum_{k=1}^L \|d_t u_h^k\|^2 \le c\tau\|\nabla u_h^0\|^2.$$

(ii) If the projection step is included, then the choice of $w_h = v_h^k$ shows that

$$\|v_h^k\|^2 + \frac{1}{2\tau}\big(\|\nabla(u_h^{k-1}+\tau v_h^k)\|^2 - \|\nabla u_h^{k-1}\|^2\big) + \frac{\tau}{2}\|v_h^k\|^2 = 0.$$

Since $\mathscr{T}_h$ is weakly acute and $u_h^k(z) = F\big(\widetilde{u}_h^k(z)\big)$ for all $z \in \mathscr{N}_h$ with the Lipschitz continuous mapping $F(s) = s/|s|$ for $|s| \ge 1$ and $F(s) = s$ otherwise, Proposition 3.2 implies as in the proof of Proposition 7.2 that

$$\|\nabla u_h^k\| \le \|\nabla[u_h^{k-1}+\tau v_h^k]\|.$$

With this, a summation over $k = 1, 2, \dots, L$, and a multiplication by τ, the previous identity implies the asserted stability estimate. To prove the estimate for the difference $v_h^k - d_t u_h^k$, let $z \in \mathscr{N}_h$. Then

$$\tau\big(d_t u_h^k(z) - v_h^k(z)\big) = u_h^k(z) - \big(u_h^{k-1}(z)+\tau v_h^k(z)\big) = \frac{\widetilde{u}_h^k(z)}{|\widetilde{u}_h^k(z)|} - \widetilde{u}_h^k(z).$$

With the identity

$$\Big|s - \frac{s}{|s|}\Big| = \Big|\frac{s}{|s|}\Big|\,\big||s|-1\big| = \big||s|-1\big|$$

for every $s \in \mathbb{R}^m$, it follows that

$$\tau\big|d_t u_h^k(z) - v_h^k(z)\big| = \big||\widetilde{u}_h^k(z)|-1\big| = \big||u_h^{k-1}(z)+\tau v_h^k(z)|-1\big|.$$

The relations $u_h^{k-1}(z)\cdot v_h^k(z) = 0$ and $|u_h^{k-1}(z)| = 1$ and the estimate $(1+s^2)^{1/2} \le 1 + s^2/2$ imply that

$$|u_h^{k-1}(z)+\tau v_h^k(z)| = \big(1+\tau^2|v_h^k(z)|^2\big)^{1/2} \le 1 + \tau^2|v_h^k(z)|^2/2.$$

A combination of the estimates and a summation over $z \in \mathscr{N}_h$ yield

$$\sum_{z\in\mathscr{N}_h} h_z^d|d_t u_h^k(z) - v_h^k(z)| \le \tau\sum_{z\in\mathscr{N}_h} h_z^d|v_h^k(z)|^2.$$

Norm equivalences and the stability result imply the asserted estimate. □

Remark 7.12 Under the conditions of Proposition 7.6 we have the local energy decay property $\|\nabla u_h^k\| \le \|\nabla u_h^{k-1}\|$ for all $k \ge 1$.

The stability estimates provide a priori bounds for the numerical approximations which allow us to pass to the limits for appropriate interpolants. Given the iterates $(u_h^k)_{k=0,\dots,K}$ of Algorithm 7.4 we define the interpolants $\widehat{u}_{h,\tau} : [0,T]\times\Omega \to \mathbb{R}^m$, $u_{h,\tau}^{\pm} : [0,T]\times\Omega \to \mathbb{R}^m$ and $v_{h,\tau}^{-} : [0,T]\times\Omega \to \mathbb{R}^m$ for $t \in (t_{k-1}, t_k)$ with $t_k = k\tau$ and $x \in \Omega$ by

$$\begin{aligned}
\widehat{u}_{h,\tau}(t,x) &= \frac{t_k - t}{\tau} u_h^{k-1}(x) + \frac{t - t_{k-1}}{\tau} u_h^k(x),\\
u_{h,\tau}^{-}(t,x) &= u_h^{k-1}(x), \quad u_{h,\tau}^{+}(t,x) = u_h^k(x),\\
v_{h,\tau}^{+}(t,x) &= v_h^k(x).
\end{aligned}$$

For ease of presentation, we again restrict the presentation to the case $m = 3$.

Theorem 7.7 (Convergence) *Suppose that $\Gamma_{\mathrm{D}} = \emptyset$, $u_h^0 \to u_0$ in $H^1(\Omega;\mathbb{R}^3)$ as $h \to 0$, and that $\mathcal{T}_h$ is weakly acute for every $h > 0$ if the projection step is carried out. Then every accumulation point of the sequence $(u_{h,\tau}^{+})_{h,\tau>0}$ in $L^\infty([0,T]; H^1(\Omega;\mathbb{R}^3))$ as $(h,\tau) \to 0$ is a weak solution of the harmonic map heat flow.*

Proof Step 1: *Selection of a weak limit.* The stability bounds of Proposition 7.6 imply that the sequences $(u_{h,\tau}^{+})_{h,\tau>0}$ and $(v_{h,\tau}^{+})_{h,\tau>0}$ are uniformly bounded in the spaces $L^\infty([0,T]; H^1(\Omega;\mathbb{R}^3))$ and $L^2([0,T]; L^2(\Omega;\mathbb{R}^3))$, respectively, so that after the extraction of a subsequence which is not relabeled, we have the existence of $u \in L^\infty([0,T]; H^1(\Omega;\mathbb{R}^3))$ and $v \in L^2([0,T]; L^2(\Omega;\mathbb{R}^3))$ with

$$\begin{aligned}
u_{h,\tau}^{\pm} &\rightharpoonup^* u \quad \text{in } L^\infty([0,T]; H^1(\Omega;\mathbb{R}^3)),\\
v_{h,\tau}^{+} &\rightharpoonup v \quad \text{in } L^2([0,T]; L^2(\Omega;\mathbb{R}^3))
\end{aligned}$$

as $(h,\tau) \to 0$. Since $v_{h,\tau}^{+} - \partial_t \widehat{u}_{h,\tau} \to 0$ in $L^2([0,T]; L^1(\Omega;\mathbb{R}^3))$ as $\tau \to 0$ we deduce that $u \in H^1([0,T]; L^2(\Omega;\mathbb{R}^3))$ and $v = \partial_t u$.

Step 2: *Verification of the energy law.* From the stability bounds we have for almost every $T' \in [0,T]$ up to a subsequence that $\nabla u_{h,\tau}^{+}(T',\cdot) \rightharpoonup \nabla u(T',\cdot)$. The weak lower semicontinuity of norms induced by inner products shows that

$$\frac{1}{2}\|\nabla u(T')\|^2 + \int_0^{T'} \|\partial_t u\|^2 \,\mathrm{d}t \le \frac{1}{2}\|\nabla u_0\|^2$$

for almost every $T' \in [0,T]$.

Step 3: *Unit-length constraint.* An interpolation estimate and $D^2 u_{h,\tau}|_R = 0$ for all elements $R \in \mathcal{T}_h$ yield for every $t \in [0,T]$ that

$$\begin{aligned}\big\|\mathscr{I}_h\big[|u_{h,\tau}^+|^2\big]-|u_{h,\tau}^+|^2\big\|_{L^1(R)} &\le c|R|^{1/2}h_R^2\big\|D^2|u_{h,\tau}^+|^2\big\|_{L^2(R)}\\ &\le c|R|^{1/2}h_R^2\|\nabla u_{h,\tau}^+\|_{L^4(R)}^2\\ &= ch_R^2|R|\big|\nabla u_{h,\tau}|_R\big|^2 \le ch_R^2\|\nabla u_{h,\tau}^+\|_{L^2(R)}^2.\end{aligned}$$

In the case of no projection we have

$$\||\mathscr{I}_h[u_{h,\tau}^+(t,\cdot)|^2]-1\|_{L^1(\Omega)} \le c\tau,$$

while for the scheme including the projection step we have, cf. the proof of Proposition 7.1,

$$\||u_{h,\tau}^+(t,\cdot)|^2-1\| \le ch\|\nabla u_h^+(t,\cdot)\|.$$

The triangle inequality yields that $|u_{h,\tau}^+|\to 1$ in $L^1([0,T]\times\Omega)$ in both cases, i.e., that $|u(t,x)|=1$ for almost every $(t,x)\in[0,T]\times\Omega$.
Step 4: *Attainment of initial data.* The weak continuity of the trace operator and $u_h^0\to u^0$ in $L^2(\Omega;\mathbb{R}^3)$ as $h\to 0$ prove $u(0,\cdot)=u_0$.
Step 5: *Passage to the limit in the equation.* It remains to show that the function u solves the partial differential equation. For this, we choose $\varphi\in L^2([0,T];C^\infty(\overline{\Omega};\mathbb{R}^3))$ and define $w^{(h,\tau)}=u_{h,\tau}^-\times\varphi$ and

$$w_{h,\tau}=\mathscr{I}_h\big[u_{h,\tau}^-\times\varphi\big].$$

For this function we have $u_{h,\tau}^-(t,z)\cdot w_{h,\tau}(t,z)=0$ for almost every $t\in[0,T]$ and every $z\in\mathscr{N}_h$. Moreover, we have using $D^2u_{h,\tau}|_R=0$ for all elements $R\in\mathscr{T}_h$ that

$$\begin{aligned}\|\nabla(w^{(h,\tau)}-w_{h,\tau})\|_{L^2(R)} &\le ch_R\|D^2[u_{h,\tau}^-\times\varphi]\|_{L^2(R)}\\ &\le h_R\big(\|\nabla u_{h,\tau}\|_{L^2(R)}\|\nabla\varphi\|_{L^2(R)}+\|u_{h,\tau}\|_{L^2(R)}\|D^2\varphi\|_{L^2(R)}\big).\end{aligned}$$

A summation over $R\in\mathscr{T}_h$ shows that $w_{h,\tau}-w^{h,\tau}\to 0$ in $L^\infty([0,T];H^1(\Omega;\mathbb{R}^3))$ as $(h,\tau)\to 0$. The equation of Algorithm 7.4 yields

$$(v_{h,\tau}^+,w_{h,\tau})+(\nabla[u_{h,\tau}^-+\tau v_{h,\tau}^+],\nabla w_{h,\tau})=0$$

for almost every $t\in[0,T]$. Due to Lemma 7.6 we have that $\tau^{1/2}v_{h,\tau}^+$ is uniformly bounded in $L^2([0,T];H^1(\Omega;\mathbb{R}^3))$ and hence the term

$$\int_0^T(\tau(\nabla v_{h,\tau}^+,\nabla w_{h,\tau})\,\mathrm{d}t \le \tau^{1/2}\Big(\int_0^T\tau\|\nabla v_{h,\tau}^+\|^2\,\mathrm{d}t\Big)^{1/2}\Big(\int_0^T\|\nabla w_{h,\tau}^+\|^2\,\mathrm{d}t\Big)^{1/2}$$

converges to 0 as $(h,\tau)\to 0$. We write

$$\int_0^T (\nabla u_{h,\tau}^-, \nabla w_{h,\tau})\,\mathrm{d}t = \int_0^T (\nabla u_{h,\tau}^-, \nabla w^{(h,\tau)})\,\mathrm{d}t + \int_0^T (\nabla u_{h,\tau}^-, \nabla[w_{h,\tau} - w^{(h,\tau)}])\,\mathrm{d}t$$

and note that the second term on the right-hand side converges to 0 as $(h, \tau) \to 0$, while for the first term on the right-hand side we have

$$\begin{aligned}\int_0^T (\nabla u_{h,\tau}^-, \nabla w^{(h,\tau)})\,\mathrm{d}t &= \int_0^T \sum_{\ell=1}^d (\partial_\ell u_{h,\tau}^-, \partial_\ell[u_{h,\tau}^- \times \varphi])\,\mathrm{d}t \\ &= \int_0^T \sum_{\ell=1}^d (\partial_\ell u_{h,\tau}^-, u \times \partial_\ell\varphi)\,\mathrm{d}t \\ &\quad + \int_0^T \sum_{\ell=1}^d (\partial_\ell u_{h,\tau}^-, [u_{h,\tau}^- - u] \times \partial_\ell\varphi)\,\mathrm{d}t.\end{aligned}$$

This implies that for $(h, \tau) \to 0$, we have

$$\int_0^T (\nabla[u_{h,\tau}^- + \tau v_{h,\tau}^+], \nabla w_{h,\tau})\,\mathrm{d}t \to \int_0^T (\nabla u, \nabla[u \times \varphi])\,\mathrm{d}t.$$

Finally, we verify that

$$\begin{aligned}\int_0^T (v_h^+, w_{h,\tau})\,\mathrm{d}t &= \int_0^T (v_h^+, u \times \varphi) + (v_h^+, [u_h - u] \times \varphi) + (v_h^+, w_{h,\tau} - w^{(h,\tau)}\,\mathrm{d}t \\ &\to \int_0^T (\partial_t u, u \times \varphi)\,\mathrm{d}t\end{aligned}$$

as $(h, \tau) \to 0$. Altogether we have proved that u satisfies

$$\int_0^T (\partial_t u, u \times \varphi) + (\nabla u, \nabla[u \times \varphi])\,\mathrm{d}t = 0$$

for all $\varphi \in L^2([0, T]; C^\infty(\overline{\Omega}; \mathbb{R}^3))$. Choosing $\varphi(t, x) = \rho(t)w(x)$ we deduce that

$$(\partial_t u, u \times w) + (\nabla u, \nabla[u \times w])\,\mathrm{d}t = 0$$

for all $w \in C^\infty(\overline{\Omega}; \mathbb{R}^3)$. A density argument proves that this is satisfied for every $w \in H^1(\Omega; \mathbb{R}^3) \cap L^\infty(\Omega; \mathbb{R}^3)$. For $\phi \in H^1(\Omega; \mathbb{R}^3) \cap L^\infty(\Omega; \mathbb{R}^3)$ and $w = u \times \phi$, we verify with the identities $u \times (u \times \phi) = (u \cdot \phi)u - \phi$ and $\partial_t u \cdot u = 0$ that

$$-(\partial_t u, \phi) + (\nabla u, \nabla[u \times (u \times \phi)]) = 0$$

for almost every $t \in [0, T]$. According to Proposition 7.5 this implies that u is a weak solution of the harmonic map heat flow. □

7.3.3 Constraint Preservation

The third characterization of solutions of the harmonic map heat flow in Proposition 7.5 reads in the strong form that

$$\partial_t u + u \times (u \times \Delta u) = 0;$$

this reveals a symplectic structure and implies that the L^2-flow of harmonic maps is constraint preserving, i.e., if $|u_0(x)| = 1$ for almost every $x \in \Omega$, then we have $|u(t, x)| = 1$ for almost every $(t, x) \in [0, T] \times \Omega$. We consider the case $\Gamma_{\mathrm{D}} = \emptyset$ for ease of presentation.

Lemma 7.3 (Constraint preservation) *Let* $u \in L^\infty([0, T]; H^1(\Omega; \mathbb{R}^3))$ *satisfy* $\partial_t u \in L^2([0, T]; L^2(\Omega; \mathbb{R}^3))$ *and* $u(0, \cdot) = u_0$ *with* u_0 *such that* $|u_0(x)| = 1$ *for almost every* $x \in \Omega$. *Assume that*

$$(\partial_t u, \phi) + (\nabla u, \nabla[u \times (u \times \phi)]) = 0$$

for almost every $t \in [0, T]$ *and every* $\phi \in H^1(\Omega; \mathbb{R}^3) \cap L^\infty(\Omega; \mathbb{R}^3)$. *Then we have* $|u(t, x)| = 1$ *for almost every* $(t, x) \in [0, T] \times \Omega$.

Proof Let $\rho \in C^\infty(\mathbb{R}^n)$ be a nonnegative function with $\|\rho\|_{L^1(B_1(0))} = 1$ and $\operatorname{supp} \rho \subset B_1(0)$. Given $\varepsilon > 0$, set $\rho_\varepsilon(x) = \rho(x/\varepsilon)$ for $x \in \Omega$. For $x_0 \in \Omega$ the choice of $\phi = \rho_\varepsilon(\cdot - x_0)u$ implies that

$$\frac{d}{dt}\frac{1}{2}(|u(t, \cdot)|^2 * \rho_\varepsilon)(x_0) = (\partial_t u, \rho_\varepsilon u) = 0,$$

i.e., $(|u(T', \cdot)|^2 * \rho_\varepsilon)(x_0) = (|u_0(\cdot)|^2 * \rho_\varepsilon)(x_0)$ for every $T' \in [0, T]$. Noting that $(|u(t, \cdot)|^2 * \rho_\varepsilon)(x_0) \to |u(t, x_0)|$ as $\varepsilon \to 0$ implies the assertion. □

The lemma motivates the development of numerical schemes that preserve the length-constraint in a discrete sense. For the Crank–Nicolson type discretization of the strong form

$$d_t u^k + u^{k-1/2} \times (u^{k-1/2} \times \Delta u^{k-1/2}) = 0$$

we observe that testing with $u^{k-1/2} = (u^k + u^{k-1})/2$ formally yields the length-preservation property

$$d_t |u^k|^2 = 0.$$

To obtain this property for a fully discrete scheme, reduced integration has to be incorporated. We define the discrete Laplacian $\widetilde{\Delta}_h u_h \in \mathscr{S}^1(\mathscr{T}_h)^3$ of a function $u_h \in \mathscr{S}^1(\mathscr{T}_h)^3$ by

$$(\widetilde{\Delta}_h u_h, v_h)_h = -(\nabla u_h, \nabla v_h)$$

for all $v_h \in \mathscr{S}^1(\mathscr{T}_h)^3$.

Algorithm 7.5 (*Constraint-preserving iteration*) Let $u_h^0 \in \mathscr{S}^1(\mathscr{T}_h)^3$ with $|u_h^0(z)| = 1$ for all $z \in \mathscr{N}_h$ and $\tau > 0$ and define the sequence $(u_h^k)_{k=0,\dots,K} \subset \mathscr{S}^1(\mathscr{T}_h)^3$ such that

$$(d_t u_h^k, \phi_h)_h + (u_h^{k-1/2} \times [u_h^{k-1/2} \times \widetilde{\Delta}_h u_h^{k-1/2}], \phi_h)_h = 0$$

for all $\phi_h \in \mathscr{S}^1(\mathscr{T}_h)^3$.

To establish the well-posedness of the algorithm, we note that a corollary of Brouwer's fixed-point theorem states that if $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ is continuous with $\Phi(s) \cdot s \geq 0$ for all $s \in \mathbb{R}^n$ with $|s| \geq R$ for some $R \in \mathbb{R}$, then there exists $s^* \in \mathbb{R}^n$ with $|s^*| \leq R$ and $\Phi(s^*) = 0$.

Proposition 7.7 (Stability and constraint preservation) *There exists a sequence $(u_h^k)_{k=0,\dots,K} \subset \mathscr{S}^1(\mathscr{T}_h)^3$ that solves the scheme of Algorithm7.5. We have $|u_h^k(z)| = 1$ for $k = 0, 1, \dots, K$ and*

$$\frac{1}{2}\|\nabla u_h^L\|^2 + \tau \sum_{k=1}^{L} \|d_t u_h^k\|^2 \leq \frac{1}{2}\|\nabla u_h^0\|^2.$$

Proof Let $k \geq 1$ and define $\Phi_h : \mathscr{S}^1(\mathscr{T}_h)^3 \to \mathscr{S}^1(\mathscr{T}_h)^3$ by

$$\Phi_h(v_h) = \frac{2}{\tau}(v_h - u_h^{k-1}) + \mathscr{I}_h\big[v_h \times (v_h \times \widetilde{\Delta}_h v_h)\big].$$

The function Φ_h is continuous and the Cauchy–Schwarz inequality, employing that $(\mathscr{I}_h w_h, v_h)_h = (w_h, v_h)_h$ for all $v_h, w_h \in \mathscr{S}^1(\mathscr{T}_h)^3$, proves that

$$(\Phi_h(v_h), v_h)_h = \frac{2}{\tau}(v_h - u_h^{k-1}, v_h)_h \geq \frac{1}{\tau}\|v_h\|_h^2 - \frac{1}{\tau}\|u_h^{k-1}\|_h^2,$$

i.e., $(\Phi_h(v_h), v_h)_h \geq 0$ for all $v_h \in \mathscr{S}^1(\mathscr{T}_h)^3$ with $\|v_h\|_h \geq \|u_h^{k-1}\|_h$. Brouwer's fixed-point theorem thus implies that there exists $r_h^k \in \mathscr{S}^1(\mathscr{T}_h)^3$ with $\Phi_h(r_h^k) = 0$ or equivalently that $u_h^k = 2r_h^k - u_h^{k-1}$ solves

$$0 = d_t u_h^k + \mathscr{I}_h\big[u_h^{k-1/2} \times (u_h^{k-1/2} \times \widetilde{\Delta}_h u_h^{k-1/2})\big],$$

i.e.,

$$(d_t u_h^k, w_h)_h + (u_h^{k-1/2} \times [u_h^{k-1/2} \times \widetilde{\Delta}_h u_h^{k-1/2}], w_h)_h = 0$$

for all $w_h \in \mathscr{S}^1(\mathscr{T}_h)^3$. For $z \in \mathscr{N}_h$ and the function $w_h = [u_h^{k-1/2}(z)]\varphi_z$, the properties of the discrete inner product imply that

$$\begin{aligned}\beta_z d_t |u_h^k(z)|^2 &= \beta_z d_t u_h^k(z) \cdot u_h^{k-1/2}(z) = (d_t u_h^k, \varphi_z u_h^{k-1/2})_h \\ &= (u_h^{k-1/2} \times [u_h^{k-1/2} \times \widetilde{\Delta}_h u_h^{k-1/2}], \varphi_z u_h^{k-1/2})_h \\ &= \beta_z\big(u_h^{k-1/2}(z) \times [u_h^{k-1/2}(z) \times \widetilde{\Delta}_h u_h^{k-1/2}(z)]\big) \cdot u_h^{k-1/2}(z) = 0,\end{aligned}$$

i.e., $|u_h^k(z)| = |u_h^{k-1}(z)|$, and inductively the assumption $|u_h^0(z)| = 1$ implies $|u_h^k(z)| = 1$. For $w_h = \widetilde{\Delta}_h u_h^{k-1/2}$, we obtain

$$\begin{aligned}&d_t \|\nabla u_h^k\|_h^2 + \big\|u_h^{k-1/2} \times \widetilde{\Delta}_h u_h^{k-1/2}\big\|_h^2 \\ &\quad= (\nabla d_t u_h^k, \nabla u_h^{k-1/2}) - (u_h^{k-1/2} \times [u_h^{k-1/2} \times \widetilde{\Delta}_h u_h^{k-1/2}], \widetilde{\Delta} u_h^{k-1/2})_h \\ &\quad= -(d_t u_h^k, \widetilde{\Delta}_h u_h^{k-1/2}) - (u_h^{k-1/2} \times [u_h^{k-1/2} \times \widetilde{\Delta}_h u_h^{k-1/2}], \widetilde{\Delta} u_h^{k-1/2})_h = 0.\end{aligned}$$

The choice of $w_h = d_t u_h^k$ shows that

$$\begin{aligned}\|d_t u_h^k\|_h^2 &= -(u_h^{k-1/2} \times \widetilde{\Delta}_h u_h^{k-1/2}, u_h^{k-1/2} \times d_t u_h^k)_h \\ &\le \|u_h^{k-1/2} \times \widetilde{\Delta}_h u_h^{k-1/2}\|_h \|u_h^{k-1/2} \times d_t u_h^k\|_h\end{aligned}$$

and with $|u_h^{k-1/2}(z)| \le 1$ for every $z \in \mathscr{N}_h$, we deduce $\|d_t u_h^k\| \le \|u_h^{k-1/2} \times \widetilde{\Delta}_h u_h^{k-1/2}\|_h$. A combination of the last two estimates, multiplication by τ, and a summation over $k = 1, 2, \dots, L$ thus prove the asserted bound. □

Remarks 7.13 (i) The stability bound implies unconditional convergence to a weak solution of the harmonic map heat flow.
(ii) The existence of the iterates in Algorithm 7.5 was established by Brouwer's fixed point theorem which is nonconstructive and in fact the iterates may not be uniquely defined. If $\tau \le c h_{\min}^2$, the following linear iteration is constraint-preserving and converges to the uniquely defined function $u_h^{k-1/2}$. Set $r_h^0 = u_h^{k-1}$ and define the sequence $(r_h^\ell)_{\ell=0,1,\dots} \subset \mathscr{S}^1(\mathscr{T}_h)^3$ via

$$\frac{2}{\tau}(r_h^\ell, \phi_h)_h + (r_h^\ell \times [r_h^{\ell-1} \times \widetilde{\Delta}_h r_h^{\ell-1}], \phi_h)_h = \frac{2}{\tau}(u_h^{k-1}, \phi_h)_h$$

for all $\phi_h \in \mathscr{S}^1(\mathscr{T}_h)^3$.

7.3.4 Approximation of Wave Maps

Wave maps are solutions of the wave equation subject to a pointwise unit-length constraint. They solve the partial differential equation

$$\partial_t^2 u - \Delta u = \lambda u$$

in $[0, T] \times \Omega$ with a Lagrange multiplier $\lambda : [0, T] \times \Omega \to \mathbb{R}$ associated to the constraint $|u(t, x)| = 1$ for almost every $(t, x) \in [0, T] \times \Omega$ and subject to the boundary condition $\partial_n u = 0$ on $[0, T] \times \partial\Omega$ and the initial conditions $u(0, \cdot) = u_0$ and $\partial_t u(0, \cdot) = u_1$. Qualitatively, similar partial differential equations arise in general relativity and particle physics. Wave maps may also be regarded as harmonic maps on $[0, T] \times \Omega$ for the Dirichlet energy defined with the Minkowski metric on the $\mathbb{R}^{1+d}$ time-space domain, i.e., they are stationary for

$$I_g(u) = \frac{1}{2} \int_0^T \int_\Omega |Du|_g^2 \,\mathrm{d}t\,\mathrm{d}x$$

with $Du = (\partial_t u, \nabla u)$ and $|v|_g^2 = -v_0^2 + v_1^2 + \cdots v_d^2$ for $v \in \mathbb{R}^{d+1}$. An important feature of solutions for the wave map equation is the energy conservation property that the mapping

$$t \mapsto I\big(u(t, \cdot), \partial_t u(t, \cdot)\big) = \frac{1}{2} \int_\Omega |\partial_t u(t, \cdot)|^2 \,\mathrm{d}x + \frac{1}{2} \int_\Omega |\nabla u(t, \cdot)|^2 \,\mathrm{d}x$$

is constant as a function of $t \in [0, T]$.

Definition 7.4 Given $u_0 \in H^1(\Omega; \mathbb{R}^m)$ and $u_1 \in L^2(\Omega; \mathbb{R}^m)$, a *wave map* is a function $u : [0, T] \times \Omega \to \mathbb{R}^m$ such that
(a) $u \in H^1([0, T]; L^2(\Omega; \mathbb{R}^m)) \cap L^2([0, T]; H^1(\Omega; \mathbb{R}^m))$,
(b) $|u(t, x)| = 1$ for almost every $(t, x) \in [0, T] \times \Omega$,
(c) for all $w \in C_c^\infty([0, T); C^\infty(\overline{\Omega}; \mathbb{R}^m))$ with $u(t, x) \cdot w(t, x) = 0$ for almost every $(t, x) \in [0, T] \times \Omega$, we have

$$-\int_0^T (\partial_t u, \partial_t w) + (\nabla u, \nabla w) \,\mathrm{d}t = (u_1, w(0)),$$

(d) the initial data u_0 is attained continuously by u as $t \to 0$ in $H^1(\Omega; \mathbb{R}^m)$,
(e) for almost every $T' \in [0, T]$, we have

$$I\big(u(T', \cdot), \partial_t u(T', \cdot)\big) \le I(u_0, u_1).$$

The algorithm for approximating wave maps is a modification of Algorithm 7.4 for the approximation of the harmonic map heat flow. The sets $\mathscr{A}_h$ and $\mathscr{F}_h[u_h^{k-1}]$ are defined as above.

Algorithm 7.6 (*Wave map approximation*) Let $u_h^0, v_h^0 \in \mathscr{S}^1(\mathscr{T}_h)^m$ with $|u_h^0(z)| = 1$ for all $z \in \mathscr{N}_h$ and $\tau > 0$ and define the sequence $(u_h^k)_{k=0,\dots,K} \subset \mathscr{S}^1(\mathscr{T}_h)^m$ for $K = \lceil T/\tau \rceil$ by computing $v_h^k \in \mathscr{F}_h[u_h^{k-1}]$ such that

$$(d_t v_h^k, w_h) + (\nabla[u_h^{k-1} + \tau v_h^k], \nabla w_h) = 0$$

for all $w_h \in \mathscr{F}_h[u_h^{k-1}]$ and setting with $\widetilde{u}_h = u_h^{k-1} + \tau v_h^k$

$$u_h^k = \widetilde{u}_h^k \qquad \text{or} \qquad u_h^k = \sum_{z \in \mathscr{N}_h} \frac{\widetilde{u}_h^k(z)}{|\widetilde{u}_h^k(z)|} \varphi_z.$$

We have the following stability result.

Proposition 7.8 (Stability) (i) *If no projection is carried out, then* $v_h^k = d_t u_h^k$ *for* $k = 1, 2, \dots, K$ *and for* $L = 1, 2, \dots, K$*, we have*

$$\frac{1}{2}\|v_h^L\|^2 + \frac{1}{2}\|\nabla u_h^L\|^2 + \frac{\tau^2}{2}\sum_{k=1}^L \big(\|d_t v_h^k\|^2 + \|\nabla v_h^k\|^2\big) = \frac{1}{2}\|v_h^0\|^2 + \frac{1}{2}\|\nabla u_h^0\|^2,$$
$$\big\|\mathscr{I}_h\big[|u_h^L|^2\big] - 1\big\|_{L^1(\Omega)} \le c_0 \tau.$$

(ii) *If a projection is carried out in every step of the algorithm and if* $\mathscr{T}_h$ *is weakly acute, then we have* $|u_h^k(z)| = 1$ *for* $k = 0, 1, \dots, K$ *and all* $z \in \mathscr{N}_h$*, and for* $L = 1, 2, \dots, K$ *that*

$$\frac{1}{2}\|v_h^L\|^2 + \frac{1}{2}\|\nabla u_h^L\|^2 + \frac{\tau^2}{2}\sum_{k=1}^L \big(\|d_t v_h^k\|^2 + \|\nabla v_h^k\|^2\big) \le \frac{1}{2}\|v_h^0\|^2 + \frac{1}{2}\|\nabla u_h^0\|^2,$$
$$\|v_h^L - d_t u_h^L\|_{L^1(\Omega)} \le c_0 \tau.$$

```
function wave_maps(d,red)
[c4n,n4e,Db,Nb] = triang_cube(d); c4n = c4n-.5;
T = 10;
tau = 2^(-red)/4; K = ceil(T/tau);
for j = 1:red
    [c4n,n4e,Db,Nb] = red_refine(c4n,n4e,Db,Nb);
end
nC = size(c4n,1);
[s,m,¬,¬] = fe_matrices(c4n,n4e);
SSS = sparse(3*nC,3*nC); MMM = sparse(3*nC,3*nC);
for k = 1 : 3
    idx = k:3:3*nC; SSS(idx,idx) = s; MMM(idx,idx) = m;
end
u = zeros(3*nC,1); v = zeros(3*nC,1);
for j = 1:nC
    u(3*j-[2,1,0]) = u_0(c4n(j,:));
    v(3*j-[2,1,0]) = v_0(c4n(j,:));
end
for k = 1:K
    B = sparse(nC,3*nC);
    for j = 1:nC
        B(j,3*j-[2,1,0]) = u(3*j-[2,1,0]);
    end
    X = [MMM+tau^2*SSS,B';B,sparse(nC,nC)];
    b = [MMM*v-tau*SSS*u;zeros(nC,1)];
    x = X\b;
    v = x(1:3*nC);
    tu = u+tau*v;
    for j = 1:nC
        u(3*j-[2,1,0]) = tu(3*j-[2,1,0])/norm(tu(3*j-[2,1,0]));
    end
    show_p1_field(c4n,u); axis(.5*[-1,1,-1,1,-1,1]);
    view(30,18); drawnow; pause(.05)
end

function val = u_0(x)
d = size(x,2);
x = [x,zeros(1,3-d)];
r = norm(x); a = max(0,1-2*r)^4;
val = [2*a*x(1:2),a^2-r^2 ]/(a^2+r^2);

function val = v_0(x)
val = [0,0,0];
```

Fig. 7.8 MATLAB realization of Algorithm 7.6 for the approximation of wave maps

Proof The choice of $w_h = v_h^k$ yields

$$\frac{d_t}{2}\|v_h^k\|^2 + \frac{\tau}{2}\|d_t v_h^k\|^2 + \frac{1}{2\tau}\left(\|\nabla[u_h^{k-1} + \tau v_h^k]\|^2 - \|\nabla u_h^{k-1}\|^2\right) + \frac{\tau}{2}\|\nabla v_h^k\|^2 = 0.$$

In the case of no projection, we have $u_h^{k-1} + \tau v_h^k = u_h^k$, and a summation over $k = 1, 2, \ldots, L$ implies the stability bound. If u_h^k is obtained through a projection, then it follows as in the proof of Proposition 7.6 that $\|\nabla u_h^k\| \le \|\nabla[u_h^{k-1} + \tau v_h^k]\|$, and again a summation over $k = 1, 2, \ldots, L$ implies the stability bound. The other estimates follow as in the proof of Proposition 7.6. □

Remark 7.14 The stability bounds imply the convergence of approximations to a wave map.

Figure 7.8 displays a MATLAB realization of Algorithm 7.6 that is based on the implementation of Algorithm 7.1.

References

1. Alouges, F.: A new algorithm for computing liquid crystal stable configurations: the harmonic mapping case. SIAM J. Numer. Anal. **34**(5), 1708–1726 (1997). http://dx.doi.org/10.1137/S0036142994264249
2. Alouges, F.: A new finite element scheme for Landau-Lifchitz equations. Discret. Contin. Dyn. Syst. Ser. S **1**(2), 187–196 (2008). http://dx.doi.org/10.3934/dcdss.2008.1.187
3. Bartels, S.: Stability and convergence of finite-element approximation schemes for harmonic maps. SIAM J. Numer. Anal. **43**(1), 220–238 (2005). http://dx.doi.org/10.1137/040606594
4. Bartels, S.: Semi-implicit approximation of wave maps into smooth or convex surfaces. SIAM J. Numer. Anal. **47**(5), 3486–3506 (2009). http://dx.doi.org/10.1137/080731475
5. Bartels, S.: Numerical analysis of a finite element scheme for the approximation of harmonic maps into surfaces. Math. Comput. **79**(271), 1263–1301 (2010). http://dx.doi.org/10.1090/S0025-5718-09-02300-X
6. Bartels, S.: Projection-free approximation of geometrically constrained partial differential equations. Math. Comput. (2013). To appear
7. Bartels, S., Prohl, A.: Constraint preserving implicit finite element discretization of harmonic map flow into spheres. Math. Comput. **76**(260), 1847–1859 (2007). http://dx.doi.org/10.1090/S0025-5718-07-02026-1
8. Chang, K.C., Ding, W.Y., Ye, R.: Finite-time blow-up of the heat flow of harmonic maps from surfaces. J. Differ. Geom. **36**(2), 507–515 (1992). http://projecteuclid.org/getRecord?id=euclid.jdg/1214448751
9. Hélein, F.: Harmonic Maps, Conservation Laws and Moving Frames. Cambridge Tracts in Mathematics, vol. 150, 2nd edn. Cambridge University Press, Cambridge (2002)
10. Lin, S.Y., Luskin, M.: Relaxation methods for liquid crystal problems. SIAM J. Numer. Anal. **26**(6), 1310–1324 (1989). http://dx.doi.org/10.1137/0726076
11. Rivière, T.: Conservation laws for conformally invariant variational problems. Invent. Math. **168**(1), 1–22 (2007). http://dx.doi.org/10.1007/s00222-006-0023-0
12. Struwe, M.: Variational Methods, 4th edn. Springer, Berlin (2008)

Chapter 8
Bending Problems

8.1 Mathematical Modeling

Bending describes the deformation of thin objects under small forces. Typically, the object is neither stretched nor sheared, but large deformations occur. A simple example is the deformation of a sheet of paper that is clamped on part of its boundary and subject to a force such as gravity. Since curvatures are important to describe such a behavior, the related mathematical models involve higher-order derivatives. We discuss the derivation of such models and their properties. For further details we refer to the textbooks [5, 6] and the seminal paper [10].

8.1.1 Bending Models

We consider a Lipschitz domain $\omega \subset \mathbb{R}^2$ representing the region occupied by a thin plate, a body force $f = (f_1, f_2, f_3)^\top : \omega \to \mathbb{R}^3$ acting on it, and clamped boundary conditions on the nonempty closed subset $\gamma_{\rm D} \subset \partial\omega$ that prescribe the displacement by a function $u_{\rm D}$ and the rotation by a mapping $\Phi_{\rm D}$ on $\gamma_{\rm D}$.

Definition 8.1 The *nonlinear Kirchhoff model* seeks a deformation $u : \omega \to \mathbb{R}^3$ that minimizes the functional

$$I^{\rm Ki}(u) = \frac{1}{2}\int_\omega |D^2 u|^2 \,{\rm d}x - \int_\omega f \cdot u \,{\rm d}x,$$

subject to the *isometry constraint* $(\nabla u)^\top \nabla u = I_2$ and the boundary conditions $u|_{\gamma_{\rm D}} = u_{\rm D}$ and $\nabla u|_{\gamma_{\rm D}} = \Phi_{\rm D}$.

The isometry constraint reflects the fact that pure bending theories do not allow for a shearing or stretching of the plate. This limits the class of boundary conditions that lead to nonempty sets of admissible deformations. In particular, the function $\Phi_{\rm D}$

S. Bartels, *Numerical Methods for Nonlinear Partial Differential Equations*,
Springer Series in Computational Mathematics 47,
DOI 10.1007/978-3-319-13797-1_8

prescribes the normal, of the deformed surface on γ_{D}. The model sets no limitations on the size of the deformation, but does not prohibit self-penetrations, i.e., it does not enforce the surface parametrized by u be embedded. We will show below that the isometry constraint allows us to replace the Frobenius norm of the Hessian by the Euclidean norm of the Laplacian, i.e., $|D^2u| = |\Delta u|$, and that these expressions coincide with the modulus of the mean curvature. For small displacements

$$\phi = u - [\mathrm{id}_2, 0]^\top,$$

i.e., if $|\nabla\phi| \ll 1$, the isometry constraint can be omitted and it suffices to consider the vertical component $w = u_3$ of the deformation. Typical large deformation and small displacement situations are depicted in Fig. 8.1.

Definition 8.2 The *linear Kichhoff model* seeks a vertical displacement $w : \omega \to \mathbb{R}$ that minimizes the functional

$$I^{\mathrm{Ki}'}(w) = \frac{1}{2}\int_\omega |D^2w|^2 \,\mathrm{d}x - \int_\omega f_3 w \,\mathrm{d}x$$

subject to the boundary conditions $w|_{\gamma_{\mathrm{D}}} = 0$ and $\nabla w|_{\gamma_{\mathrm{D}}} = 0$, i.e., w belongs to the set $H^2_{\mathrm{D}}(\omega) = \{v \in H^2(\Omega) : v|_{\gamma_{\mathrm{D}}} = 0,\ \nabla v|_{\gamma_{\mathrm{D}}} = 0\}$.

The linear Kirchhoff model is closely related to a model in which no second-order derivatives occur. It may be regarded as an approximation of the linear Kirchhoff model in which small shearing effects may occur. Mathematically, the second order derivatives are replaced by an additional variable and the difference is penalized with a penalty parameter, which may be regarded as a small artificial plate thickness. Notice that the symmetric gradient of a gradient is the Hessian, i.e., $\varepsilon(\nabla w) = D^2w$.

Definition 8.3 The *linear Reissner–Mindlin model* seeks for given $t > 0$ a vertical displacement $w : \omega \to \mathbb{R}$ and a rotation $\theta : \omega \to \mathbb{R}^3$ that minimize the functional

$$I^{\mathrm{RM}}(w, \theta) = \frac{t^{-2}}{2}\int_\omega |\theta - \nabla w|^2 \,\mathrm{d}x + \frac{1}{2}\int_\omega |\varepsilon(\theta)|^2 \,\mathrm{d}x - \int_\omega f_3 w \,\mathrm{d}x,$$

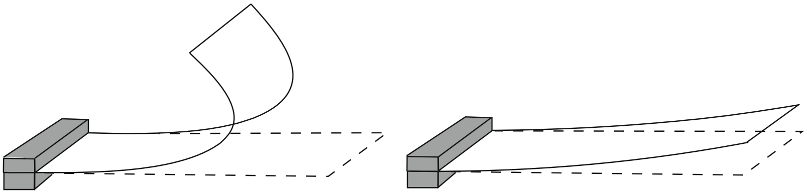


Fig. 8.1 Large isometric deformation of a thin clamped plate (*left*) and small displacement described by a linear model (*right*)

where $\varepsilon(\theta) = [(\nabla\theta)^\top + (\nabla\theta)]/2$, subject to the boundary conditions $w|_{\gamma_\mathrm{D}} = 0$ and $\theta|_{\gamma_\mathrm{D}} = 0$.

A solution u of the nonlinear Kirchhoff model defines an open surface in $\mathbb{R}^3$ that is parametrized by the deformation u. Since this surface is isometric to ω, we have that the Gaussian curvature K vanishes, i.e., that the local length and angle relations are preserved under the deformation. The mean curvature is given by $H^2 = |D^2u|^2$ and this identity establishes a relation to a bending model that is used to describe the deformation of fluid membranes such as cell surfaces. Here, the considered surfaces are closed. The justification of the model is less clear than in the case of solids. In particular, fluid membranes can undergo large shearing effects that are not seen by its description as a surface.

Definition 8.4 The *Willmore model* seeks a closed surface $\mathscr{M} \subset \mathbb{R}^3$ that minimizes the functional

$$I^{\mathrm{Wi}}(\mathscr{M}) = \frac{1}{2}\int_{\mathscr{M}} H^2 \,\mathrm{d}s - \int_{\mathscr{M}} K \,\mathrm{d}s,$$

subject to constraints that the surface area of $\mathscr{M}$ or that the volume enclosed by $\mathscr{M}$ be prescribed.

The integral over the Gaussian curvature is a topological invariant and can be neglected if a minimizer is sought in a fixed topology class. If the surface area and the enclosed volume are prescribed, then the model is referred to as the *Helfrich model*.

8.1.2 Relations to Hyperelasticity

In three-dimensional hyperelasticity, pure bending is characterized by a cubic scaling of the energy with respect to the plate thickness t, i.e., that

$$I_t(u_t) = \int_{\Omega_t} W(\nabla u_t)\,\mathrm{d}x - \int_{\Omega_t} f_t \cdot u_t \,\mathrm{d}x \sim t^3$$

for the optimal deformations $u_t \in H^1(\Omega_t;\mathbb{R}^3)$ as $t \to 0$ for $\Omega_t = \omega \times (-t/2, t/2) \subset \mathbb{R}^3$, such that $u_t|_{\Gamma_\mathrm{D}} = \mathrm{id}$ on $\Gamma_\mathrm{D} = \gamma_\mathrm{D} \times (-t/2, t/2)$. This motivates considering the rescaled energy functionals $\widehat{I_t} = t^{-3} I_t$ and investigating the limiting behavior for $t \to 0$ in the framework of Γ-convergence. We let ∇' denote the gradient with respect to the first two variables $x' = (x_1, x_2)$. The corresponding three-dimensional objects are denoted $\nabla = (\nabla', \partial_3)$ and $x = (x', x_3)$.

Theorem 8.1 (Dimension reduction [10]) *Let*

$$W(F) = \mathrm{dist}^2\big(F, SO(3)\big)$$

for all $F \in \mathbb{R}^{3\times 3}$ *and* $SO(3) = \{F \in \mathbb{R}^{3\times 3} : F^\top F = I_3, \det F = 1\}$. *Set* $\widehat{f}_t(x', \widehat{x}_3) = t^{-2} f_t(x', t\widehat{x}_3)$ *and assume* $\widehat{f}_t \to f$ *in* $L^2(\Omega_1; \mathbb{R}^3)$ *and that* f *is independent of* $\widehat{x}_3 \in (-1, 1)$. *Let* $(u_t)_{t>0}$ *be a sequence of minimizers for the sequence of functionals* $(I_t)_{t>0}$, *i.e.,* $u_t \in H^1(\Omega_t; \mathbb{R}^3)$ *with* $u_t|_{\Gamma_{\mathrm{D}}} = \mathrm{id}_{\Gamma_{\mathrm{D}}}$. *Then the rescaled functions* $\widehat{u}(x', \widehat{x}_3) = u(x', t\widehat{x}_3)$ *converge in* $H^1(\Omega_1; \mathbb{R}^3)$ *to a function* $u \in H^1(\Omega_1; \mathbb{R}^3)$. *This function is independent of* $\widehat{x}_3$, *defines a parametrized surface with the first fundamental form* $g = (\nabla' u)^\top (\nabla' u) = I_2$ *in* Ω_1, *and satisfies* $u \in H^2(\Omega_1; \mathbb{R}^3)$. *Moreover, it has the boundary values* $u|_{\gamma_{\mathrm{D}}} = [\mathrm{id}, 0]^\top$ *and* $\nabla' u|_{\gamma_{\mathrm{D}}} = [I_2, 0]^\top$ *and minimizes*

$$I^{\mathrm{Ki}}(u) = \frac{1}{12} \int_\omega |h|^2 \,\mathrm{d}x' - \int_\omega f \cdot u \,\mathrm{d}x',$$

with the normal $b = \partial_1 u \times \partial_2 u$ *and the second fundamental form* $h = -(\nabla' b)^\top (\nabla' u)$, *in functions* $v \in H^1(\Omega_1; \mathbb{R}^3)$, *that are independent of* $\widehat{x}_3$, *satisfy* $(\nabla' v)^\top (\nabla' v) = I_2$ *in* Ω_1, *and have the same boundary conditions as* u. *Conversely, every such minimizer* u *of* I^{Ki} *is the limit of a sequence of rescaled minimizers of* I_t *and the minimal energies converge to* $I^{\mathrm{Ki}}(u)$.

Remarks 8.1 (i) We will show below that $|h| = |D^2 u|$ for the Frobenius norms of the second fundamental form and the Hessian of u.
(ii) The result also holds for isotropic, frame-indifferent energy densities $W \in C^2(\mathbb{R}^{n\times n})$ with $W(I_3) = 0$, and $W(F) \geq \mathrm{dist}^2\big(F, SO(3)\big)$, cf. [10].

For a heuristic justification of the result, we follow [7] and consider the rescaled energy functional

$$\widehat{I}_t(u) = t^{-3} \int_{\Omega_t} W(\nabla u) \,\mathrm{d}x$$

with W given by

$$W(F) = \mathrm{dist}^2\big(F, SO(3)\big) = \min_{Q \in SO(3)} |F - Q|^2.$$

We assume that the optimal deformation $u_t = u$ is of the form

$$u(x', x_3) = v(x') + x_3 b(x')$$

with t-independent vector fields $v, b : \omega \to \mathbb{R}^3$ and b is normal to the surface parametrized by v, i.e., $\partial_\ell v(x') \cdot b(x') = 0$ for $\ell = 1, 2$. This means that v is the deformation of the middle surface ω and the segments normal to ω are mapped to straight lines that are normal to the deformed surface, cf. the right plot of Fig. 8.2. We have

$$\nabla u = [\nabla' v, b] + [x_3 \nabla' b, 0].$$

For matrices $F \in \mathbb{R}^{3\times 3}$ in a neighborhood of $SO(3)$, we use the approximation

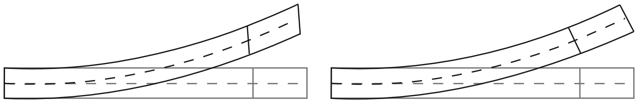


Fig. 8.2 Normal segments are mapped to straight line segments under the Reissner–Mindlin hypotheses (*left*); the Kirchhoff–Love hypotheses require that the deformed segments be normal to the deformed middle surface (*right*)

$$W(F) = \operatorname{dist}^2\big(F, SO(3)\big) \approx \frac{1}{4}|F^\top F - I_3|^2.$$

For a proof of this relation consider $F = P + \varepsilon G$, where $P = \pi_{SO(3)}(F)$ is the nearest-neighbor projection of F onto $SO(3)$ and G is normal to $SO(3)$ at P. We may assume that $P = I_3$, which implies that G is symmetric. Then $\operatorname{dist}^2\big(F, SO(3)\big) = \varepsilon^2|G|^2$ and $|F^\top F - I_3|^2 = \varepsilon^2|G+G^\top|^2 + \mathscr{O}(\varepsilon^3) = 4\varepsilon^2|G|^2 + \mathscr{O}(\varepsilon^3)$. Since $\widehat{I_t}(u) = t^{-3}I_t(u) \le C$ and t is small, we expect that $W(\nabla u)$ is small, i.e., that ∇u is close to $SO(3)$ so that

$$\widehat{I_t}(u) \approx \frac{t^{-3}}{4}\int\limits_{\Omega_t} \big|(\nabla u)^\top \nabla u - I_3\big|^2 \,\mathrm{d}x.$$

Noting $(\nabla' v)^\top \nabla' b = (\nabla' b)^\top \nabla' v$, we have

$$(\nabla u)^\top \nabla u = \begin{bmatrix} (\nabla' v)^\top \nabla' v & 0 \\ 0 & |b|^2 \end{bmatrix} + x_3 \begin{bmatrix} 2(\nabla' b)^\top \nabla' v & (\nabla' b)^\top b \\ b^\top \nabla' b & 0 \end{bmatrix} + x_3^2 \begin{bmatrix} (\nabla' b)^\top \nabla' b & 0 \\ 0 & 0 \end{bmatrix}.$$

With the abbreviations

$$\widehat{g}_t = t^{-1}\big((\nabla' v)^\top \nabla' v - I_2\big), \quad h = -(\nabla' v)^\top \nabla' b, \quad k = (\nabla' b)^\top b,$$

we obtain

$$\begin{aligned} \widehat{I_t}(u) &\approx \frac{t^{-3}}{4}\int\limits_{\Omega_t} \left| \begin{bmatrix} t\widehat{g}_t & 0 \\ 0 & |b|^2 - 1 \end{bmatrix} + x_3 \begin{bmatrix} -2h & (\nabla' b)^\top b \\ b^\top (\nabla' b) & 0 \end{bmatrix} + x_3^2 \begin{bmatrix} k & 0 \\ 0 & 0 \end{bmatrix} \right|^2 \mathrm{d}x \\ &= \frac{t^{-3}}{4}\int\limits_{\Omega_t} \left| \begin{bmatrix} t\widehat{g}_t - 2x_3 h + x_3^2 k & (\nabla' b)^\top b \\ b^\top (\nabla' b) & |b|^2 - 1 \end{bmatrix} \right|^2 \mathrm{d}x. \end{aligned}$$

To guarantee that this expression is bounded t-independently, we need to impose the condition $|b|^2 = 1$, and with the resulting identity $b^\top \nabla' b = 0$, we deduce that

$$\widehat{I_t}(u) \approx \frac{t^{-3}}{4}\int\limits_{\Omega_t} \big|t\widehat{g}_t - 2x_3 h + x_3^2 k\big|^2 \,\mathrm{d}x.$$

By carrying out the integration with respect to x_3, we obtain

$$\widehat{I}_t(u) \approx \frac{1}{4}\int\limits_{\omega} |\widehat{g}_t|^2 + \frac{1}{3}|h|^2 + \frac{t^2}{5\cdot 2^4}|k|^2 + \frac{t}{6}\widehat{g}_t : k \,\mathrm{d}x'.$$

Again, to obtain a t-independent limit, we need that $\widehat{g}_t = 0$. Neglecting the term involving the factor t^2, this leads to the reduced, t-independent functional

$$\widehat{I}_t(u) = \frac{1}{12}\int\limits_{\omega} |h|^2 \,\mathrm{d}x',$$

subject to the pointwise constraint $(\nabla' v)^\top \nabla' v = I_2$. We finally remark that for forces described by functions f_t that are independent of x_3 and such that $t^{-2} f_t \to f$ in $L^2(\omega;\mathbb{R}^3)$ as $t \to 0$, we find with the assumed expansion $u(x) = v(x') + x_3 b(x')$ that

$$\begin{aligned} t^{-3}\int\limits_{\Omega_t} f_t \cdot u \,\mathrm{d}x &= t^{-3}\int\limits_{\Omega_t} f_t \cdot v \,\mathrm{d}x + t^{-3}\int\limits_{-t/2}^{t/2}\int\limits_{\omega} x_3 b \cdot f_t \,\mathrm{d}x' \,\mathrm{d}x_3 \\ &= t^{-2}\int\limits_{\Omega_t} f_t \cdot v \,\mathrm{d}x \to \int\limits_{\omega} f \cdot v \,\mathrm{d}x' \end{aligned}$$

as $t \to 0$.

8.1.3 Relations to Linear Elasticity

Linear elasticity employs a *geometric linearization* defined through the symmetric gradient

$$\varepsilon(\phi) = \frac{1}{2}\big((\nabla\phi)^\top + \nabla\phi\big) \approx \frac{1}{2}\big((\nabla u)^\top \nabla u - I_3\big)$$

for small displacements $\phi = u - \mathrm{id}_3 : \Omega \to \mathbb{R}^3$ with $\Omega \subset \mathbb{R}^3$. The energy density W is approximated by the quadratic expression

$$W(\nabla u) \approx \frac{1}{2} D^2 W(I_3)[\nabla\phi, \nabla\phi] = \frac{1}{2} D^2 W(I_3)[\varepsilon(\phi), \varepsilon(\phi)],$$

provided W is isotropic and frame-indifferent, using that $W(I_3) = 0$, and $D\widetilde{W}(I_3) = 0$. For homogeneous materials it follows that with the Lamé constants λ, μ we have for every symmetric matrix $E \in \mathbb{R}^{3\times 3}$ with $\mathbb{C} = D^2 W(I_3)$ that

$$\mathbb{C}E = 2\mu E + \lambda(\operatorname{tr} E)I_3.$$

The related minimization problem looks for $\phi : \Omega \to \mathbb{R}^3$ to be minimal for the *Navier–Lamé functional*

$$I^{\mathrm{NL}}(\phi) = \frac{1}{2}\int_\Omega \mathbb{C}\varepsilon(\phi) : \varepsilon(\phi)\,\mathrm{d}x - \int_\Omega \widehat{f}\cdot\phi\,\mathrm{d}x,$$

subject to $\phi|_{\Gamma_\mathrm{D}} = 0$. For thin plates $\Omega_t = \omega \times (-t/2, t/2)$ with Dirichlet boundary $\Gamma_\mathrm{D} = \gamma_\mathrm{D} \times (-t/2, t/2)$ for $\gamma_\mathrm{D} \subset \partial\omega$, often the following assumptions are made to obtain a dimensionally reduced model. The different assumptions are illustrated in Fig. 8.2.

Assumption 8.1 (*Reissner–Mindlin hypotheses*) (1) Points on the middle surface are only displaced in the vertical direction, i.e., $\phi_1(x', 0) = \phi_2(x', 0) = 0$ for all $x' \in \omega$.
(2) The vertical displacement does not depend on x_3, i.e., $\phi_3(x', x_3) = w(x')$.
(3) Segments that are normal to the middle surface are linearly deformed, i.e., $\phi(x', x_3) = \phi(x', 0) - x_3\widehat{\theta}(x')$ for all $(x', x_3) \in \Omega_t$.

The assumption implies that the minimizer for I^{NL} is given by

$$\phi(x', x_3) = \begin{bmatrix} -x_3\theta(x') \\ w(x') \end{bmatrix}$$

with the rotation $\theta : \omega \to \mathbb{R}^2$ and the vertical displacement $w : \omega \to \mathbb{R}$.

Assumption 8.2 (*Kirchhoff–Love hypotheses*) In addition to the Reissner–Mindlin hypotheses, assume that segments that are normal to the middle surface are mapped linearly and isometrically to segments that are normal to the deformed middle surface, i.e., $\phi(x', x_3) = \phi(x', 0) - x_3\widehat{\theta}(x')$ for all $(x', x_3) \in \Omega_t$ with

$$\widehat{\theta}(x', 0) = (1 + |\nabla' w|^2)^{-1/2}\begin{bmatrix} \nabla' w \\ 0 \end{bmatrix} \approx \begin{bmatrix} \nabla' w \\ 0 \end{bmatrix}.$$

Note that ϕ is the displacement, so that the third component of the normal vector $\widehat{\theta}$ disappears. The additional assumption implies that the solution of the linearly elastic problem is given by

$$\phi(x', x_3) = \begin{bmatrix} -x_3\nabla' w(x') \\ w(x') \end{bmatrix}$$

for the vertical displacement $w : \omega \to \mathbb{R}$.

Proposition 8.1 (Linear bending) *Assume that* f_t *is independent of* x_3 *and set* $f_3 = t^{-2}f_{t,3}$. *Suppose that* $\mathbb{C}E = E$ *for all symmetric matrices* $E \in \mathbb{R}^{3\times 3}$. *Let* $\phi \in H^1_\mathrm{D}(\Omega_t; \mathbb{R}^3)$ *be the minimizer of the three-dimensional elasticity functional* I^{NL} *with*

$\Omega = \Omega_t$ and $\widehat{f} = f_t$. Up to a change of constants we have:
(i) Under the Reissner–Mindlin hypotheses the pair $(w, \theta) \in H^1_{\mathrm{D}}(\omega) \times H^1_{\mathrm{D}}(\omega; \mathbb{R}^2)$ that specifies ϕ solves the linear Reissner–Mindlin model.
(ii) Under the Kirchhoff–Love hypotheses the function $w \in H^2_{\mathrm{D}}(\omega)$ that specifies ϕ solves the linear Kirchhoff model.

Proof In the case of the Reissner–Mindlin hypotheses we have

$$\varepsilon'(\phi) = \frac{1}{2}\begin{bmatrix} -x_3\nabla'\theta & -\theta \\ (\nabla' w)^\top & 0 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} -x_3(\nabla'\theta)^\top & \nabla' w \\ -\theta^\top & 0 \end{bmatrix} = \begin{bmatrix} -x_3\varepsilon'(\theta) & (\nabla' w - \theta)/2 \\ (\nabla' w - \theta)^\top/2 & 0 \end{bmatrix}.$$

Therefore, due to the assumption $\mathbb{C}E = E$,

$$\mathbb{C}\varepsilon'(\phi) : \varepsilon'(\phi) = x_3^2|\varepsilon'(\theta)|^2 + \frac{1}{2}|\nabla' w - \theta|^2.$$

An integration over $\Omega_t = \omega \times (-t/2, t/2)$ shows that

$$\frac{1}{2}\int_{\Omega_t} \mathbb{C}\varepsilon'(\varphi) : \varepsilon'(\varphi)\,\mathrm{d}x = \frac{t^3}{24}\int_{\omega} |\varepsilon'(\theta)|^2\,\mathrm{d}x' + \frac{t}{4}\int_{\omega} |\nabla' w - \theta|^2\,\mathrm{d}x'.$$

Since f_t is independent of x_3, we have

$$\int_{\Omega_t} f_t \cdot \varphi\,\mathrm{d}x = \int_{\omega}\int_{-t/2}^{t/2} (-x_3)\theta \cdot f_{t,12}\,\mathrm{d}x_3\,\mathrm{d}x' + \int_{\omega}\int_{-t/2}^{t/2} w f_{t,3}\,\mathrm{d}x_3\,\mathrm{d}x' = t\int_{\omega} f_{t,3} w\,\mathrm{d}x'.$$

Hence,

$$t^{-3} I^{\mathrm{NL}}(\varphi) = \frac{1}{24}\int_{\omega} |\varepsilon(\theta)|^2\,\mathrm{d}x' + \frac{t^{-2}}{4}\int_{\omega} |\nabla w - \theta|^2\,\mathrm{d}x' - \int_{\omega} f_3 w\,\mathrm{d}x'.$$

For the Kirchhoff hypothesis, this simplifies to $I^{\mathrm{Ki}'}$ due to the identities $\nabla' w = \theta$ and $\varepsilon'(\nabla' w) = \nabla'\nabla' w$. □

Remark 8.2 If $\mathbb{C}E = 2\mu E + \lambda(\operatorname{tr} E)I_3$ is considered then the assumption that for $\sigma = \mathbb{C}\varepsilon(\phi)$ we have $\sigma_{33} = 0$ has to be included.

8.1.4 Properties of Isometries

Given a surface $\mathscr{M}$ parametrized by $u : \omega \to \mathbb{R}^3$ the *first and second fundamental forms* $g, h : \omega \to \mathbb{R}^{2\times 2}$ are given by

$$g = (\partial_i u \cdot \partial_j u)_{1 \le i,j \le 2} = (\nabla u)^\top \nabla u,$$
$$h = -(\partial_i b \cdot \partial_j u)_{1 \le i,j \le 2} = -(\nabla b)^\top \nabla u = b^\top D^2 u,$$

where $b = \partial_1 u \times \partial_2 u / |\partial_1 u \times \partial_2 u|$ is a unit normal to $\mathscr{M}$. The parametrization is assumed to be an immersion, so that the tangent vectors $\partial_1 u$ and $\partial_2 u$ are linearly independent everywhere in ω. The first and second fundamental form are interpreted as bilinear forms on the tangent space $T\mathscr{M}$ in terms of the coefficients of the family of bases $\big(\partial_1 u(x), \partial_2 u(x)\big)_{x \in \omega}$. It follows that g is a symmetric and positive definite matrix for every $x \in \omega$ that defines a metric on the tangent space of $\mathscr{M}$. The *Gauss and mean curvature* are the determinant and the trace of the *Weingarten map*

$$s = -hg^{-1}$$

and given by

$$K = \det s = \frac{\det h}{\det g}, \quad H = \operatorname{tr} s = -\frac{h : \det' g}{\det g},$$

respectively. The Weingarten map measures variations of the normal b and is interpreted as a linear mapping on the tangent space. The second fundamental form is the bilinear form associated with s. We refer the reader to Sect. 8.4 for a detailed discussion.

Definition 8.5 The parametrization $u : \omega \to \mathbb{R}^3$ is called *isometry* if $g(x) = I_2$ for every $x \in \omega$.

Proposition 8.2 *Suppose that $u : \omega \to \mathbb{R}^3$ is a C^2-isometry. Then $\partial_i \partial_j u \cdot \partial_k u = 0$, $K = 0$, and*

$$|D^2 u| = |\Delta u| = |h| = |H|,$$

where $|\cdot|$ denotes the Frobenius norm on the respective spaces.

Proof We first note that for $1 \le i, j \le 2$, we have $0 = \partial_i(\partial_j u \cdot \partial_j u) = 2\partial_i \partial_j u \cdot \partial_j u$. To show that we also have $\partial_i^2 u \cdot \partial_j u = 0$ for $i \neq j$, we note $0 = \partial_i(\partial_i \cdot \partial_j u) = \partial_i^2 u \cdot \partial_j u + \partial_i u \cdot \partial_i \partial_j u$, i.e., $\partial_i^2 u \cdot \partial_j u = -\partial_i u \cdot \partial_i \partial_j u = 0$. Hence, we have

$$\partial_i \partial_j u \cdot \partial_k u = 0$$

for $i, j, k = 1, 2$, i.e., the Christoffel symbols of the second kind vanish. As a consequence of Gauss' theorem, cf. Lemma 8.3, we have $K = 0$. Moreover, we deduce that $-\Delta u = \beta b$ and since $(-\Delta u) \cdot b = \operatorname{tr}(-h) = H$, we have $\beta = H$. The vectors $(\partial_1 u, \partial_2 u, b)$ form an orthonormal basis of $\mathbb{R}^3$ for every $x \in \omega$, so that $|\partial_i \partial_j u| = |\partial_i \partial_j u \cdot b|$ and hence

$$|D^2 u|^2 = \sum_{i,j=1}^{2} |\partial_i \partial_j u \cdot b|^2 = |h|^2.$$

Moreover, we have

$$|h|^2 = |s|^2 = (\operatorname{tr} s)^2 - 2 \det s = H^2 - 2K = H^2,$$

which proves the assertion. □

Remark 8.3 Since isometries in $H^2(\omega; \mathbb{R}^3)$ can be approximated by isometries in $C^2(\overline{\omega}; \mathbb{R}^3)$ in the norm of $H^2(\omega; \mathbb{R}^3)$, the results of the proposition also hold for isometries $u \in H^2(\omega; \mathbb{R}^3)$, cf. [12].

8.2 Approximaton of Linear Bending Models

We discuss in this section numerical methods for the approximation of the linear Kirchhoff and the linear Reissner–Mindlin model. Finite element methods for dimensionally reduced models have to be carefully developed to avoid so-called *locking effects*. This describes the phenomenon that deformations obtained by numerical computation are too small in comparison to the true deformation. In particular, *membrane locking* is the inability of a finite element method to capture bending effects without stretching while *shear locking* refers to the problem that a finite element method is too stiff to describe certain in-plane deformations due to the occurrence of a small parameter. Another effect that occurs in the description of thin elastic structures is the *Babuška paradox* that states that if a domain is approximated by polygons, then the numerical solutions may fail to converge to the correct solution. We follow closely the presentation of [5] and refer the reader to [4] for further aspects.

8.2.1 Discrete Kirchhoff Triangles

To avoid an H^2-conforming finite element method for the linear Kirchhoff model, we employ a nonconforming discretization that is based on the construction of a discrete gradient operator

$$\nabla_h : W_h \to \Theta_h$$

with H^1-conforming finite element spaces $W_h \subset H^1(\omega)$ and $\Theta_h \subset H^1(\omega; \mathbb{R}^2)$. These are for a regular triangulation $\mathscr{T}_h$ of ω defined as

$$\begin{aligned} W_h = \{w_h \in C(\overline{\omega}) : w_h|_T \in P_3^{\mathrm{red}}(T) \text{ for all } T \in \mathscr{T}_h, \\ \nabla w_h \text{ continuous at all } z \in \mathscr{N}_h\}, \\ \Theta_h = \{\theta_h \in C(\overline{\omega}) : \theta_h|_T \in P_2(T) \text{ for all } T \in \mathscr{T}_h\}. \end{aligned}$$

Here, $P_k(T)$ for every $T \in \mathscr{T}_h$ denotes the set of polynomials of total degree less or equal to $k \geq 0$ restricted to T. The superscript in P_3^{red} means that one degree of

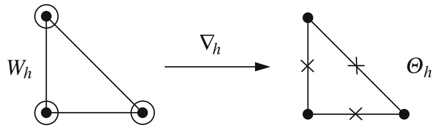


Fig. 8.3 Schematic description of the elementwise reduced cubic finite element space W_h (*left*) and the space of elementwise quadratic vector fields Θ_h (*right*)

freedom is eliminated, i.e., with the center of mass $x_T = (1/3)\sum_{z\in\mathscr{N}_h\cap T} z$ of T,

$$P_3^{\mathrm{red}}(T) = \big\{p \in P_3(T) : p(x_T) = \frac{1}{3}\sum_{z\in\mathscr{N}_h\cap T}\big[p(z) + \nabla p(z)\cdot(x_T - z)\big]\big\}.$$

The degrees of freedom in W_h are the function values and the derivatives at the vertices of the elements, cf. Fig. 8.3. For $w \in H^3(\omega)$, we define the nodal interpolant $\widetilde{\mathscr{I}}_h^3 w \in W_h$ by the conditions $\widetilde{\mathscr{I}}_h^3 w(z) = w(z)$ and $\nabla\widetilde{\mathscr{I}}_h^3 w(z) = \nabla w(z)$ for all $z \in \mathscr{N}_h$.

Definition 8.6 The *discrete gradient operator* $\nabla_h : W_h \to \Theta_h$ is for $w_h \in W_h$ the uniquely defined function $\theta_h = \nabla_h w_h \in \Theta_h$ with

$$\begin{aligned}
\theta_h(z) &= \nabla w_h(z) && \text{for all } z \in \mathscr{N}_h,\\
\theta_h(z_S)\cdot n_S &= \frac{1}{2}\big(\nabla w_h(z_S^1) + \nabla w_h(z_S^2)\big)\cdot n_S && \text{for all } S \in \mathscr{S}_h,\\
\theta_h(z_S)\cdot t_S &= \nabla w_h(z_S)\cdot t_S && \text{for all } S \in \mathscr{S}_h,
\end{aligned}$$

where, for all sides $S \in \mathscr{S}_h$, the orthonormal vectors $n_S, t_S \in \mathbb{R}^2$ are chosen such that n_S is normal to S, $z_S^1, z_S^2 \in \mathscr{N}_h$ are the endpoints of S, and $z_S = (z_S^1 + z_S^2)/2$ is the midpoint of S. For $w \in H^3(\Omega)$, we set $\nabla_h w = \nabla_h \widetilde{\mathscr{I}}_h^3 w$.

Remark 8.4 For every $S \in \mathscr{S}_h$ we have

$$\nabla_h w_h(z_S) = \frac{1}{2}\big[\big(\nabla w_h(z_S^1) + \nabla w_h(z_S^2)\big)\cdot n_S\big]n_S + \big[\nabla w_h(z_S)\cdot t_S\big]t_S.$$

The following lemma shows that ∇_h may be regarded as an interpolation operator on the space of gradients of functions in $H^3(\omega)$. We let $\gamma_{\mathrm{D}} \subset \partial\omega$ be closed and of positive surface measure and define $\gamma_{\mathrm{N}} = \partial\omega \setminus \gamma_{\mathrm{D}}$.

Lemma 8.1 (Properties of ∇_h [5]) (i) *There exists* $c_1 > 0$ *such that for all* $w_h \in W_h$ *and* $T \in \mathscr{T}_h$*, we have for* $\ell = 0, 1$ *that*

$$c_1^{-1}\|\nabla^{\ell+1} w_h\|_{L^2(T)} \le \|\nabla^{\ell}\nabla_h w_h\|_{L^2(T)} \le c_1\|\nabla^{\ell+1} w_h\|_{L^2(T)},$$

where $\nabla^1 = \nabla$ *and* $\nabla^0 = I$.
(ii) *There exists* $c_2 > 0$ *such that for all* $w \in H^3(\omega)$ *and* $T \in \mathscr{T}_h$, *we have*

$$\|\nabla_h w - \nabla w\|_{L^2(T)} + h_T \|\nabla\nabla_h w - D^2 w\|_{L^2(T)} \le c_2 h_T^2 \|D^3 w\|_{L^{(}T)}.$$

(iii) *There exists* $c_3 > 0$ *such that for all* $w_h \in W_h$ *and* $T \in \mathscr{T}_h$, *we have*

$$\|\nabla_h w_h - \nabla w_h\|_{L^2(T)} \le c_3 h_T \|D^2 w_h\|_{L^2(T)}.$$

(iv) *The mapping* $w_h \mapsto \|\nabla\nabla_h w_h\|$ *defines a norm on*

$$W_{h,\mathrm{D}} = \big\{w_h \in W_h : w_h(z) = 0,\ \nabla w_h(z) = 0 \textit{ for all } z \in \mathscr{N}_h \cap \gamma_\mathrm{D}\big\},$$

and we have $w_h|_{\gamma_\mathrm{D}} = 0$ *and* $\nabla w_h|_{\gamma_\mathrm{D}} = 0$ *for all* $w_h \in W_{h,\mathrm{D}}$.

Proof (i) Both expressions define semi-norms and we show that $\nabla^{\ell+1} w_h = 0$ if and only if $\nabla^\ell \nabla_h w_h = 0$ for all $w_h \in W_h$. Assume that $\nabla_h w_h|_T = c_T$ for some $c_T \in \mathbb{R}^2$. Then $\nabla w_h(z) = c_T$ for all $z \in \mathscr{N}_h \cap T$ and $\nabla w_h(z_S) = c_T$ for all $S \in \mathscr{S}_h \cap T$. Thus, the cubic polynomials $w_h|_S$ are affine for all $S \in \mathscr{S}_h \cap \partial T$, and also the function $w_h|_{\partial T}$ is affine. Due to the elementwise constraint in the definition of W_h, it follows that $w_h|_T$ is affine and thus $\nabla w_h = c_T$. If conversely $\nabla w_h|_T = c_T$, then also $\nabla_h w_h|_T = c_T$. Hence, the expressions $\|\nabla^{\ell+1} w_h\|_{L^2(T)}$ and $\|\nabla^\ell \nabla_h w_h\|_{L^2(T)}$ are equivalent semi-norms on $W_h|_T$ and a scaling argument proves the first assertion.
(ii) Since $\nabla_h w|_T$ is affine if $\nabla w|_T$ is affine, the Bramble–Hilbert lemma yields the interpolation estimate

$$\|\theta - \theta_h\|_{L^2(T)} + h_T \|\nabla(\theta - \theta_h)\|_{L^2(T)} \le c h_T^2 \|D^2 \theta\|_{L^2(T)}$$

for $\theta = \nabla w \in H^2(\omega)$ and $\theta_h = \nabla_h w$.
(iii) The estimate is a consequence of (ii) and the inverse estimate $\|D^3 w_h\|_{L^2(T)} \le c h_T^{-1} \|D^2 w_h\|_{L^2(T)}$.
(iv) If $w_h(z) = 0$ and $\nabla_h w_h(z) = 0$ for all $z \in \mathscr{N}_h \cap \gamma_\mathrm{D}$ then, since $w_h|_S$ is a cubic polynomial for every $S \in \mathscr{S}_h$, it follows that $w_h|_{\gamma_\mathrm{D}} = 0$ and $\nabla_h w_h|_{\gamma_\mathrm{D}} = 0$. Assume that $\|\nabla\nabla_h w_h\| = 0$. Then, since $\nabla_h w_h|_{\gamma_\mathrm{D}} = 0$ we deduce by Poincaré inequality that $\nabla_h w_h = 0$ in ω. With (i) and $w_h|_{\gamma_\mathrm{D}} = 0$ we find $w_h = 0$ in ω. □

The interpolation estimates allow us to prove the following error estimate.

Theorem 8.2 (Error estimate) *Assume that* $w \in H^2_\mathrm{D}(\omega) \cap H^3(\omega)$ *is the solution of the linear Kirchhoff model, i.e.,*

$$(D^2 w, D^2 v) = (f, v)$$

for all $v \in H^2_\mathrm{D}(\omega)$ *and let* $w_h \in W_{h,\mathrm{D}}$ *solve*

$$(\nabla\nabla_h w_h, \nabla\nabla_h v_h) = (f, v_h)$$

for all $v_h \in W_{h,\mathrm{D}}$. *Then we have*

$$\|D^2 w - \nabla\nabla_h w_h\| \le ch\|w\|_{H^3(\omega)}.$$

Proof The Lax–Milgram lemma and Lemma 8.1(iv) imply the existence of unique solutions $w \in H^2_{\mathrm{D}}(\omega)$ and $w_h \in W_{h,\mathrm{D}}$. The assumption $w \in H^3(\omega)$, the boundary condition $(D^2 w)n|_{\gamma_{\mathrm{N}}} = 0$, an integration by parts, and the identities div $D^2 = \Delta\nabla = \nabla\Delta$ show, that for all $v \in H^2_{\mathrm{D}}(\omega)$, we have

$$(f, v) = (D^2 w, D^2 v) = -(\nabla\Delta w, \nabla v)$$

and this identity holds for all $v \in H^1_{\mathrm{D}}(\omega)$. Therefore, for $v_h \in W_{h,\mathrm{D}}$ it follows that

$$\begin{aligned}(\nabla\nabla_h w, \nabla\nabla_h v_h) &= (D^2 w, \nabla\nabla_h v_h) + (\nabla[\nabla_h w - \nabla w], \nabla\nabla_h v_h)\\ &= -(\nabla\Delta w, \nabla_h v_h) + (\nabla[\nabla_h w - \nabla w], \nabla\nabla_h v_h)\\ &= -(\nabla\Delta w, \nabla v_h) - (\nabla\Delta w, [\nabla_h v_h - \nabla v_h])\\ &\quad + (\nabla[\nabla_h w - \nabla w], \nabla\nabla_h v_h).\end{aligned}$$

Recalling that $\nabla_h w = \nabla_h \widetilde{\mathscr{I}}_h^3 w$ and incorporating the discrete and continuous formulations, this yields that

$$\begin{aligned}\|\nabla\nabla_h[w - w_h]\|^2 &= (\nabla\nabla_h w, \nabla\nabla_h[w - w_h]) - (\nabla\nabla_h w_h, \nabla\nabla_h[w - w_h])\\ &= (f, \widetilde{\mathscr{I}}_h^3 w - w_h) + (\nabla\Delta w, \nabla_h[w - w_h] - \nabla[\widetilde{\mathscr{I}}_h^3 w - w_h])\\ &\quad + (\nabla[\nabla_h w - \nabla w], \nabla\nabla_h[w - w_h]) - (f, \widetilde{\mathscr{I}}_h^3 w - w_h)\\ &= (\nabla\Delta w, \nabla_h[w - w_h] - \nabla[\widetilde{\mathscr{I}}_h^3 w - w_h])\\ &\quad + (\nabla[\nabla_h w - \nabla w], \nabla\nabla_h[w - w_h]).\end{aligned}$$

For the first term on the right-hand side we have by Lemma 8.1(i) and (iii) that

$$(\nabla\Delta w, \nabla_h[w - w_h] - \nabla[\widetilde{\mathscr{I}}_h^3 w - w_h]) \le ch\|\nabla\Delta w\|\,\|\nabla\nabla_h[w - w_h]\|.$$

The second term is estimated with the help of Lemma 8.1(ii), i.e.,

$$(\nabla[\nabla_h w - \nabla w], \nabla\nabla_h[w - w_h]) \le ch\|D^3 w\|\,\|\nabla\nabla_h[w - w_h]\|$$

The combination of the last three estimates, the triangle inequality, and the bound $\|D^2 w - \nabla\nabla_h w\| \le ch\|D^3 w\|$ of Lemma 8.1(ii) prove the assertion. □

8.2.2 Realization

For the implementation of the discrete Kirchhoff triangle, we identify functions $w_h \in W_h$ and $\theta_h \in \Theta_h$ with vectors $W \in \mathbb{R}^{3L}$ and $\Theta \in \mathbb{R}^{2(L+M)}$, where $L = n_C = \#\mathscr{N}_h$ and $M = n_S = \#\mathscr{S}_h$, defined by

$$W = \begin{bmatrix} w_h(z_1) \\ \nabla w_h(z_1) \\ w_h(z_2) \\ \nabla w_h(z_2) \\ \vdots \\ w_h(z_L) \\ \nabla w_h(z_L) \end{bmatrix} = \begin{bmatrix} w_{z_1} \\ \delta w_{z_1} \\ w_{z_2} \\ \delta w_{z_2} \\ \vdots \\ w_{z_L} \\ \delta w_{z_L} \end{bmatrix}, \quad \Theta = \begin{bmatrix} \theta_h(z_1) \\ \theta_h(z_2) \\ \vdots \\ \theta_h(z_L) \\ \theta_h(z_{S_1}) - \big(\theta_h(z_{S_1}^1) + \theta_h(z_{S_1}^2)\big)/2 \\ \theta_h(z_{S_2}) - \big(\theta_h(z_{S_2}^1) + \theta_h(z_{S_2}^2)\big)/2 \\ \vdots \\ \theta_h(z_{S_M}) - \big(\theta_h(z_{S_M}^1) + \theta_h(z_{S_M}^2)\big)/2 \end{bmatrix} = \begin{bmatrix} \theta_{z_1} \\ \theta_{z_2} \\ \vdots \\ \theta_{z_L} \\ \theta_{S_1} \\ \theta_{S_2} \\ \vdots \\ \theta_{S_M} \end{bmatrix}$$

with $\mathscr{N}_h = \{z_1, z_2, \dots, z_L\}$ and $\mathscr{S}_h = \{S_1, S_2, \dots, S_M\}$. For the coefficient of θ_h related to a side $S \in \mathscr{S}_h$, we subtract half of the values of θ_h at the corresponding endpoints z_S^1 and z_S^2 since we use the hierarchical basis

$$\big(\varphi_{z_1}, \varphi_{z_2}, \dots, \varphi_{z_L}, \varphi_{S_1}, \varphi_{S_2}, \dots \varphi_{S_M}\big)$$

of the space $\mathscr{S}^2(\mathscr{T}_h) = \{v_h \in C(\overline{\omega}) : v_h|_T \in P_2(T) \text{ for all } T \in \mathscr{T}_h\}$ given by the nodal basis $(\varphi_{z_1}, \varphi_{z_2}, \dots, \varphi_{z_L})$ of $\mathscr{S}^1(\mathscr{T}_h)$ and the functions $\varphi_S = 4\varphi_{z_S^1}\varphi_{z_S^2}$ for all $S \in \mathscr{S}_h$. A straightforward calculation shows that, for a function $w_h \in P_3^{\mathrm{red}}(T)$, we have that $w_h|_S$ is cubic for every side $S \subset \partial T$ with

$$\big(\nabla w_h(z_S)\big) \cdot t_S = \frac{3}{2|S|}\big(w_h(z_S^2) - w_h(z_S^1)\big) - \frac{1}{4}\big(\nabla w_h(z_S^1) + \nabla w_h(z_S^2)\big) \cdot t_S$$

with $|S| = |z_S^2 - z_S^1|$ and $z_S^2 - z_S^1 = |S| t_S$. Since (n_S, t_S) are orthonormal vectors it follows for $\theta_h = \nabla_h w_h$ that

$$\begin{aligned} \theta_h(z_S) &= \big(\nabla w_h(z_S) \cdot t_S\big)t_S + \Big[\frac{1}{2}\big(\nabla w_h(z_S^1) + \nabla w_h(z_S^2)\big) \cdot n_S\Big]n_S \\ &= \big(\nabla w_h(z_S) \cdot t_S\big)t_S + \frac{1}{2}\big(\nabla w_h(z_S^1) + \nabla w_h(z_S^2)\big) \\ &\quad - \Big[\frac{1}{2}\big(\nabla w_h(z_S^1) + \nabla w_h(z_S^2)\big) \cdot t_S\Big]t_S \\ &= \frac{3}{2|S|}\big(w_h(z_S^2) - w_h(z_S^1)\big)t_S - \frac{3}{4}\Big[\big(\nabla w_h(z_S^1) + \nabla w_h(z_S^2)\big) \cdot t_S\Big]t_S \\ &\quad + \frac{1}{2}\big(\nabla w_h(z_S^1) + \nabla w_h(z_S^2)\big). \end{aligned}$$

Since $\theta_h(z_S^j) = \nabla w_h(z_S^j), j = 1, 2$, the corresponding coefficient is given by

$$\begin{aligned}\theta_S &= \theta_h(z_S) - \big(\theta_h(z_S^1) + \theta_h(z_S^2)\big)/2 \\ &= \frac{3}{2|S|}\big(w_h(z_S^2) - w_h(z_S^1)\big)t_S - \frac{3}{4}\big[\big(\nabla w_h(z_S^1) + \nabla w_h(z_S^2)\big) \cdot t_S\big]t_S.\end{aligned}$$

With these identifications, the discrete gradient operator can be represented by a matrix $D_h \in \mathbb{R}^{2(L+M)\times 3L}$. For a single element $T = \operatorname{conv}\{z_1, z_2, z_3\}$ with sides $S_1 = \operatorname{conv}\{z_2, z_3\}$, $S_2 = \operatorname{conv}\{z_3, z_1\}$, and $S_3 = \operatorname{conv}\{z_1, z_2\}$, we have

$$\begin{bmatrix}\theta_{z_1}\\ \theta_{z_2}\\ \theta_{z_3}\\ \theta_{S_1}\\ \theta_{S_2}\\ \theta_{S_3}\end{bmatrix} = \begin{bmatrix} 0 & I_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & I_2 \\ 0 & 0 & \widetilde{t}_{S_1} & \widetilde{T}_{S_1} & -\widetilde{t}_{S_1} & \widetilde{T}_{S_1} \\ \widetilde{t}_{S_2} & \widetilde{T}_{S_2} & 0 & 0 & -\widetilde{t}_{S_2} & \widetilde{T}_{S_2} \\ \widetilde{t}_{S_3} & \widetilde{T}_{S_3} & -\widetilde{t}_{S_3} & \widetilde{T}_{S_3} & 0 & 0 \end{bmatrix} \begin{bmatrix} w_{z_1}\\ \delta w_{z_1}\\ w_{z_2}\\ \delta w_{z_2}\\ w_{z_3}\\ \delta w_{z_3}\end{bmatrix}$$

where $\widetilde{T}_{S_\ell} = -(3/4)t_{S_\ell}t_{S_\ell}^\top$ and $\widetilde{t}_{S_\ell} = -(3/(2|S_\ell|))t_{S_\ell}$. For a simpler implementation we approximated the right-hand side using numerical integration, i.e.,

$$\int_\omega f_3 w_h \,\mathrm{d}x \approx \int_\omega \mathscr{I}_h[f_3 w_h]\,\mathrm{d}x$$

which is computed with the lumped mass matrix. Figure 8.5 displays an implementation of the approximation of the linear Kirchhoff model with the discrete Kirchhoff triangle. The $M \times 2$ field `n4s` provides an enumeration of the edges and defines their endpoints. The field `s4e` has dimension $n_E \times 3$, $n_E = \#\mathscr{T}_h$, and contains the global numbers of the sides of the elements in $\mathscr{T}_h$, where the convention that the jth edge of T is opposite to the jth node of T is used, cf. Fig. 8.4. These arrays are provided by the subroutine `sides`. The stiffness matrix of the $P2$ finite element space with respect to the hierarchical basis defined above is provided by the routine `fe_matrix_p2.m`.

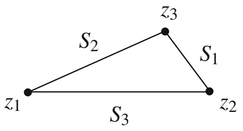


Fig. 8.4 Local enumeration of the sides of a *triangle* every side is associated to the opposite node

```
function kirchhoff_linear(red)
[c4n,n4e,Db,Nb] = triang_cube(2);
for j = 1:red
    [c4n,n4e,Db,Nb] = red_refine(c4n,n4e,Db,Nb);
end
[n4s,s4e] = sides(n4e);
nC = size(c4n,1); nS = size(n4s,1);
dNodes = unique(Db);
FNodes = setdiff(1:3*nC,[3*dNodes-2;3*dNodes-1;3*dNodes-0]);
u = zeros(3*nC,1); b = zeros(3*nC,1);
D = sparse(2*(nC+nS),3*nC);
for j = 1:nC
    D(2*j-[1,0],3*j-[1,0]) = eye(2);
end
for j = 1:nS
    t_S = (c4n(n4s(j,2),:)-c4n(n4s(j,1),:))';
    length_S = norm(t_S); t_S = t_S/length_S;
    D(2*nC+2*j-[1,0],3*n4s(j,1)-2) = -3/(2*length_S)*t_S;
    D(2*nC+2*j-[1,0],3*n4s(j,2)-2) = 3/(2*length_S)*t_S;
    D(2*nC+2*j-[1,0],3*n4s(j,1)-[1,0]) = -(3/4)*(t_S*t_S');
    D(2*nC+2*j-[1,0],3*n4s(j,2)-[1,0]) = -(3/4)*(t_S*t_S');
end
[s_p1,¬,m_lumped,vol_T] = fe_matrices(c4n,n4e);
s_p2 = fe_matrix_p2(c4n,n4e,n4s,s4e,s_p1,vol_T);
S = sparse(2*(nC+nS),2*(nC+nS));
S(1:2:2*(nC+nS),1:2:2*(nC+nS)) = s_p2;
S(2:2:2*(nC+nS),2:2:2*(nC+nS)) = s_p2;
S_dkt = D'*S*D;
b(3*(1:nC)-2) = m_lumped*f(c4n);
u(FNodes) = S_dkt(FNodes,FNodes)\b(FNodes);
show_p1(c4n,n4e,Db,Nb,u(1:3:3*nC))

function [n4s,s4e] = sides(n4e)
sides = reshape(n4e(:,[2,3,3,1,1,2])',2,[])';
[n4s,¬,sideNrs] = unique(sort(sides,2),'rows','first');
s4e = reshape(sideNrs(1:3*size(n4e,1)),3,[])';

function val = f(x)
val = ones(size(x,1),1);
```

Fig. 8.5 MATLAB routine for the approximation of the linear Kirchhoff model with Kirchhoff triangles

8.2.3 Reissner–Mindlin Plate

The linear Reissner–Mindlin model seeks a pair $(w,\theta)\in H^1_\mathrm{D}(\omega)\times H^1_\mathrm{D}(\omega;\mathbb{R}^2)$ such that

$$\big(\varepsilon(\theta),\varepsilon(\psi)\big)+t^{-2}(\theta-\nabla w,\psi-\nabla\eta)=(f,\eta)$$

for all $(\psi,\eta)\in H^1_\mathrm{D}(\omega;\mathbb{R}^2)\times H^1_\mathrm{D}(\omega)$. The corresponding strong form of the problem reads as

$$
\begin{aligned}
-\operatorname{div}\,\varepsilon(\theta)+t^{-2}(\theta-\nabla w)&=0 \text{ in } \omega, \quad \theta|_{\gamma_{\mathrm{D}}}=0, \quad \partial_n\theta|_{\gamma_{\mathrm{N}}}=0,\\
t^{-2}\operatorname{div}(\theta-\nabla w)&=f \text{ in } \omega, \quad w|_{\gamma_{\mathrm{D}}}=0, \quad (\theta-\nabla w)\cdot n|_{\gamma_{\mathrm{N}}}=0
\end{aligned}
$$

with $\gamma_{\mathrm{N}}=\partial\omega\setminus\gamma_{\mathrm{D}}$. The problem can be simplified by employing a Helmholtz decomposition of $\theta-\nabla w$. For a function $p\in H^1(\omega)$ we write

$$
\operatorname{Curl} p=(\nabla p)^{\perp}=[-\partial_2 p,\partial_1 p]^{\top}.
$$

Proposition 8.3 (Equivalent formulation) *Assume that ω is simply connected. There exist uniquely defined functions $r\in H^1_{\mathrm{D}}(\omega)$ and $p\in H^1(\omega)$ with $\int_\omega p\,\mathrm{d}x=0$ and* $\operatorname{Curl} p\cdot n|_{\gamma_{\mathrm{N}}}=0$*, such that $t^{-2}(\theta-\nabla w)=-\nabla r-\operatorname{Curl} p$. The function $r\in H^1_{\mathrm{D}}(\omega)$ satisfies*

$$
(\nabla r,\nabla\eta)=(f,\eta)
$$

for all $\eta\in H^1_{\mathrm{D}}(\omega)$. The pair (θ,p) is uniquely defined by the equations

$$
\begin{aligned}
\big(\varepsilon(\theta),\varepsilon(\psi)\big)-\quad(\operatorname{Curl} p,\psi)\quad&=(\nabla r,\psi),\\
(\theta,\operatorname{Curl} q)-t^2(\operatorname{Curl} p,\operatorname{Curl} q)&=0
\end{aligned}
$$

for all $(\psi,q)\in H^1_{\mathrm{D}}(\omega;\mathbb{R}^2)\times H^1(\omega)$ with $\operatorname{Curl} q\cdot n|_{\gamma_{\mathrm{N}}}=0$*. The function $w\in H^1_{\mathrm{D}}(\omega)$ satisfies*

$$
(\nabla w,\nabla v)=(\theta,\nabla v)+t^2(\nabla r,\nabla v)
$$

for all $v\in H^1_{\mathrm{D}}(\omega)$.

Proof Let $r\in H^1_{\mathrm{D}}(\omega)$ be the unique solution of

$$
(\nabla r,\nabla\eta)=(f,\eta)=-t^{-2}(\theta-\nabla w,\nabla\eta)
$$

for all $\eta\in H^1_{\mathrm{D}}(\omega)$. Since $F=t^{-2}(\theta-\nabla w)+\nabla r$ satisfies $\operatorname{div} F=0$ in ω and since $F\cdot n|_{\gamma_{\mathrm{N}}}=0$, there exists a uniquely defined function $p\in H^1(\omega)$ with $\int_\omega p\,\mathrm{d}x=0$, $\operatorname{Curl} p\cdot n=0$ on γ_{N}, and $F=-\operatorname{Curl} p$, cf., e.g., [11]. For all $\eta\in H^1_{\mathrm{D}}(\omega)$, we then have

$$
(\operatorname{Curl} p,\nabla\eta)=\int_{\partial\omega}\eta\,\operatorname{Curl} p\cdot n\,\mathrm{d}s=0.
$$

The equations now follow from the weak formulation of the linear Reissner–Mindlin model and the identity that defines Curl p. □

The equations derived in the proposition show that the solution of the linear Reissner-Mindlin model can be computed by successively solving three problems. The first and the third formulations that define r and w are Poisson problems, while the second one defines the pair (θ,p) through a saddle-point problem with a penalty term that is qualitatively equivalent to the Stokes problem. In particular, the inf-sup condition is satisfied and the solution operator is bounded t-independently. This implies the robust solvability of the Reissner–Mindlin model, provided that the finite element spaces used for the approximation of (θ, p) satisfy a discrete inf-sup condition. A possible choice is the so-called *mini-element*, which is the lowest order conforming polynomial element for the Stokes problem. To guarantee that a discrete Helmholtz decomposition is available, the variables r and w then need to be approximated in the nonconforming Crouzeix–Raviart finite element space, cf. [1] for related details and optimal, t-independent error estimates.

8.3 Approximation of the Nonlinear Kirchhoff Model

The linear Kirchhoff model may be regarded as a simplification of the nonlinear Kirchhoff model in the case of small displacements. We generalize in this section the finite element method based on discrete Kirchhoff triangles for the linear model to the nonlinear one that describes large bending deformations. The proposed method uses techniques developed in [3].

8.3.1 Discretization

We employ the spaces W_h and Θ_h introduced for the approximation of the linear Kirchhoff model. The fact that the gradient of a function in W_h is continuous at vertices of elements allows us to impose the isometry constraint at those points. We thus consider the minimization problem defined by

$$I_h^{\mathrm{Ki}}(u_h) = \frac{1}{2}\int_\omega |\nabla \nabla_h u_h|^2 \,\mathrm{d}x - \int_\omega f \cdot u_h \,\mathrm{d}x$$

$$\text{subject to } u_h \in \mathscr{A}_h = \big\{ v_h \in W_h^3, \quad [\nabla v_h(z)]^\top \nabla v_h(z) = I_2 \text{ for all } z \in \mathscr{N}_h, \\ v_h(z) = u_{\mathrm{D}}(z),\ \nabla v_h(z) = \Phi_{\mathrm{D}}(z) \text{ for all } z \in \mathscr{N}_h \cap \gamma_{\mathrm{D}} \big\}.$$

For the vector field $u_h \in W_h^3$, the approximate gradient $\nabla_h u_h$ is obtained by applying ∇_h to each component of u_h. We suppose that the boundary data u_{D} and Φ_{D} are compatible in the sense that for a function $\widetilde{u}_{\mathrm{D}} \in H^2(\omega; \mathbb{R}^3)$ with $(\nabla \widetilde{u}_{\mathrm{D}})^\top \nabla \widetilde{u}_{\mathrm{D}} = I_2$ in ω, we have $u_{\mathrm{D}} = \widetilde{u}_{\mathrm{D}}|_{\gamma_{\mathrm{D}}}$ and $\Phi_{\mathrm{D}} = \nabla \widetilde{u}_{\mathrm{D}}|_{\gamma_{\mathrm{D}}}$. We also assume that u_{D} and Φ_{D} can be approximated with arbitrary accuracy by nodal interpolation on γ_{D}, i.e.,

$$\big\| u_{\mathrm{D}} - \mathscr{I}_h \widetilde{u}_{\mathrm{D}}|_{\gamma_{\mathrm{D}}} \big\|_{L^2(\gamma_{\mathrm{D}})} + \big\| \Phi_{\mathrm{D}} - \mathscr{I}_h \nabla \widetilde{u}_{\mathrm{D}}|_{\gamma_{\mathrm{D}}} \big\|_{L^2(\gamma_{\mathrm{D}})} \to 0$$

as $h \to 0$. For analyzing convergence of the numerical scheme, we assume that there exists a solution of the nonlinear Kirchhoff model that is smooth or which can be approximated by smooth isometries. This assumption is not a restriction because of corresponding density results in [12].

Theorem 8.3 (Approximation) *Assume that there exists a minimizer $u \in \mathscr{A}$ with*

$$\mathscr{A} = \{v \in H^2(\omega; \mathbb{R}^3) : (\nabla v)^\top \nabla v = I_2,\ v|_{\gamma_\mathrm{D}} = u_\mathrm{D},\ \nabla v|_{\gamma_\mathrm{D}} = \Phi_\mathrm{D}\}$$

for the nonlinear Kirchhoff model which can be approximated in $H^2(\omega; \mathbb{R}^3)$ by functions $v \in \mathscr{A} \cap H^3(\omega; \mathbb{R}^3)$. For every $h > 0$ there exists a minimizer $u_h \in W_h^3$ of I_h^{Ki}. If $(u_h)_{h>0}$ is a sequence of minimizers, then $\|\nabla u_h\| \le C$, for all $h > 0$, and every accumulation point $u \in H^1(\omega; \mathbb{R}^3)$ of the sequence is a strong accumulation point, belongs to $H^2(\omega; \mathbb{R}^3)$, satisfies $(\nabla u)^\top \nabla u = I_2$ almost everywhere in ω, $u|_{\gamma_\mathrm{D}} = u_\mathrm{D}$, and $\nabla u|_{\gamma_\mathrm{D}} = \Phi_\mathrm{D}$, and is a minimizer for I^{Ki}.

Proof By Lemma 8.1 (iii) we have that $\|\nabla\nabla_h u_h\|$ is a norm and this implies that I_h^{Ki} has a minimizer. Because of the assumptions on the boundary data, it follows by Poincaré inequality and Lemma 8.1 (i) that $\|\nabla u_h\| \le C$ and $\|\nabla\nabla_h u_h\| \le C$ for all $h > 0$. Let $u \in H^1(\omega; \mathbb{R}^3)$ and $z \in H^1(\omega; \mathbb{R}^{3\times 2})$ be such that for a subsequence (which is not relabeled), we have $u_h \rightharpoonup u$ in $H^1(\omega; \mathbb{R}^3)$ and $\nabla_h u_h \rightharpoonup z$ in $H^1(\omega; \mathbb{R}^{3\times 2})$. With Lemma 8.1 we verify that $\|\nabla_h u_h - \nabla u_h\| \le ch\|\nabla\nabla_h u_h\|$ and this yields $\nabla u = z$, in particular $u \in H^2(\omega; \mathbb{R}^3)$. The attainment of the boundary data follows from continuity properties of the trace operators and the fact that

$$\|u_h - \mathscr{I}_h u_h\| + \|\nabla_h u_h - \mathscr{I}_h \nabla_h u_h\| \to 0$$

as $h \to 0$. A nodal interpolation estimate and an inverse estimate yield that for every $T \in \mathscr{T}_h$, we have

$$\begin{aligned}\big\|(\nabla u_h)^\top \nabla u_h - I_2\big\|_{L^1(T)} &\le ch_T^2 \big\|D^2\big[(\nabla u_h)^\top \nabla u_h\big]\big\|_{L^1(T)} \\ &\le ch_T^2\big(\|D^3 u_h\|_{L^2(T)}\|\nabla u_h\|_{L^2(T)} + \|D^2 u_h\|_{L^2(T)}^2\big) \\ &\le ch_T\big(\|D^2 u_h\|_{L^2(T)}\|\nabla u_h\|_{L^2(T)} + \|D^2 u_h\|_{L^2(T)}^2\big).\end{aligned}$$

A summation over all $T \in \mathscr{T}_h$ together with the fact that ∇u_h converges strongly to ∇u implies that $(\nabla u)^\top \nabla u = I_2$ almost everywhere in ω. To verify that u minimizes I^{Ki}, we first note that by weak lower semicontinuity of the L^2 norm, we have

$$\|D^2 u\| = \|\nabla z\| \le \liminf_{h\to 0} \|\nabla\nabla_h u_h\|$$

and

$$\int_\omega u_h \cdot f \,\mathrm{d}x \to \int_\omega u \cdot f \,\mathrm{d}x.$$

This proves that

$$I^{\mathrm{Ki}}(u) \le \liminf_{h\to 0} I_h^{\mathrm{Ki}}(u_h).$$

To show that the minimal energy is attained let $\widetilde{u} \in \mathscr{A}$ be a minimizing isometry for I^{Ki}. Due to the assumed approximability of $\widetilde{u}$ by smooth isometries, we may assume that $\widetilde{u} \in H^3(\omega; \mathbb{R}^3)$. We define $\widetilde{u}_h = \widetilde{\mathscr{I}}_h^3 \widetilde{u} \in \mathscr{A}_h$ and note with Lemma 8.1(ii) that

$$\|\nabla_h \widetilde{u}_h - \nabla \widetilde{u}\| + h\|\nabla\nabla_h \widetilde{u}_h - D^2\widetilde{u}\| \le ch^2\|\widetilde{u}\|_{H^3(\omega)}$$

which implies the attainment of the minimal energy. □

8.3.2 Iterative Minimization

Our iterative scheme for the practical solution of the discretized minimization problem realizes a discrete H^2-gradient flow of the energy functional with a linearization of the nodal isometry constraint about the current iterate. For this, it is important to realize that for the employed finite element space W_h, the nodal values of the discrete deformation $\big(u_h(z) : z \in \mathscr{N}_h\big)$ and its gradient $\big(\nabla u_h(z) : z \in \mathscr{N}_h\big)$ are mutually independent variables in the minimization problem.

Algorithm 8.1 (Discrete H^2-isometry-flow) *Let $\tau > 0$ and $u_h^0 \in W_h^3$ be such that*

$$\big[\nabla u_h^0(z)\big]^\top \nabla u_h^0(z) = I_2$$

for all $z \in \mathscr{N}_h$ and $u_h^0(z) = u_{\mathrm{D}}(z)$ and $\nabla_h u_h^0(z) = \Phi_{\mathrm{D}}(z)$ for all $z \in \mathscr{N}_h \cap \gamma_{\mathrm{D}}$. For $k = 1, 2, \dots$, define

$$\begin{aligned}&\mathscr{F}_h[u_h^{k-1}]\\ &\quad = \big\{w_h \in W_{h,\mathrm{D}}^3 : [\nabla w_h(z)]^\top \nabla u_h^{k-1}(z) + [\nabla u_h^{k-1}(z)]^\top \nabla w_h(z) = 0 \text{ f.a. } z \in \mathscr{N}_h\big\}\end{aligned}$$

and compute $u_h^k = u_h^{k-1} + \tau d_t u_h^k$ with $d_t u_h^k \in \mathscr{F}_h[u_h^{k-1}]$ satisfying

$$\big(\nabla\nabla_h d_t u_h^k, \nabla\nabla_h w_h\big) + \alpha\big(\nabla\nabla_h(u_h^{k-1} + \tau d_t u_h^k), \nabla\nabla_h w_h\big) = \big(f, w_h\big)$$

for all $w_h \in \mathscr{F}_h[u_h^{k-1}]$. Stop the iteration if $\|\nabla\nabla_h d_t u_h^k\| \le \varepsilon_{\mathrm{stop}}$.

The iterates $(u_h^k)_{k=0,1,\dots}$ will in general not satisfy the nodal isometry constraint exactly, but the violation is independent of the number of iterations and controlled by the step size τ.

Theorem 8.4 (Iteration) *The iterates $(u_h^k)_{k=0,1,\dots}$ of Algorithm 8.1 are well defined and satisfy*

$$I_h^{\mathrm{Ki}}(u_h^k) + \frac{\tau}{2}\|\nabla\nabla_h d_t u_h^k\|^2 \le I_h^{\mathrm{Ki}}(u_h^{k-1}).$$

Moreover, we have

$$\|\mathscr{I}_h\big[(\nabla u_h^k)^\top \nabla u_h^k\big] - I_2\|_{L^1(\omega)} \le C\tau I_h^{\mathrm{Ki}}(u_h^0).$$

Proof The existence of a unique $d_t u_h^k \in F_h[u_h^{k-1}]$ in every step of the iteration follows from the fact that the bilinear form $(v_h, w_h) \mapsto \big(\nabla\nabla_h v_h, \nabla\nabla_h w_h\big)$ defines a coercive and continuous bilinear form on $\mathscr{F}_h[u_h^{k-1}]$, cf. Lemma 8.1(iv). Upon choosing $w_h = d_t u_h^k$, we find that

$$\big\|\nabla\nabla_h d_t u_h^k\big\|^2 + \frac{1}{2} d_t \big\|\nabla\nabla_h u_h^k\big\|^2 + \frac{\tau}{2}\big\|\nabla\nabla_h d_t u_h^k\big\|^2 = \big(f, d_t u_h^k\big)$$

and this proves the energy decreasing property. Using $u_h^k = u_h^{k-1} + \tau d_t u_h^k$, we have

$$\begin{aligned}\big(\nabla u_h^k\big)^\top \nabla u_h^k = \big(\nabla u_h^{k-1}\big)^\top \nabla u_h^{k-1} + \tau\big(\nabla d_t u_h^k\big)^\top \nabla u_h^{k-1} \\ + \tau\big(\nabla u_h^{k-1}\big)^\top \nabla d_t u_h^k + \tau^2\big(\nabla d_t u_h^k\big)^\top \nabla d_t u_h^k.\end{aligned}$$

Since $d_t u_h^k \in F_h[u_h^{k-1}]$, the sum of the second and third term on the right-hand side vanishes at every $z \in \mathscr{N}_h$ and an inductive argument, together with the assumptions on u_h^0, leads to

$$\big|\big[\nabla u_h^L(z)\big]^\top \nabla u_h^L(z) - I_2\big| \le \tau^2 \sum_{k=1}^{L} \big|\nabla d_t u_h^k(z)\big|^2.$$

A discrete norm equivalence and a local inverse inequality imply the assertion. □

8.3.3 Realization

The implementation of Algorithm 8.1 is based on the realization of the discrete Kirchhoff triangle for the linear problem. We also employ quadrature to discretize the forcing term which we assume to act only in the vertical direction. This implies that only the nodal values $\big(u_h(z) : z \in \mathscr{N}_h\big)$ and $\big(\nabla u_h(z) : z \in \mathscr{N}_h\big)$ are needed for the implementation, in particular, no evaluation of u_h in the interior of elements in $\mathscr{T}_h$ is required. If S_2 is the stiffness matrix related to piecewise quadratic vector fields with six components, D realizes the operator $\nabla_h : W_h^3 \to \Theta_h^3$, and B_{k-1} encodes the constraints and boundary conditions defined in the space $\mathscr{F}_h[u_h^{k-1}]$, then one step of the discrete gradient flow leads to the linear system of equations

$$\begin{bmatrix}(1+\alpha\tau)D^\top S_2 D & B_{k-1}^\top \\ B_{k-1} & 0\end{bmatrix}\begin{bmatrix}d_t U^k \\ \Lambda\end{bmatrix} = \begin{bmatrix}-\alpha D^\top S_2 D\, U^{k-1} + \tau F \\ 0\end{bmatrix}.$$

```
function kirchhoff_nonlinear(red)
[c4n,n4e,Db,Nb] = triang_strip(10);
alpha = 1; tau = 2ˆ(-red)/10;
for j = 1:red
    [c4n,n4e,Db,Nb] = red_refine(c4n,n4e,Db,Nb);
end
nC = size(c4n,1);
dNodes = unique(Db); DNodes = [3*dNodes-2;3*dNodes-1;3*dNodes-0];
FNodes = setdiff(1:9*nC,[0*nC+DNodes;3*nC+DNodes;6*nC+DNodes]);
S_dkt = fe_matrix_dkt(c4n,n4e);
[¬,¬,m_lumped] = fe_matrices(c4n,n4e);
Z = sparse(3*nC,3*nC);
SSS = [S_dkt,Z,Z;Z,S_dkt,Z;Z,Z,S_dkt];
SSS_free = SSS(FNodes,FNodes);
u = u_moebius(c4n);
dt_u = zeros(9*nC,1);
bbb = zeros(9*nC,1);
bbb(6*nC+(1:3:3*nC)) = m_lumped*f3(c4n);
corr = 1; eps_stop = 1e-2;
while corr > eps_stop;
    B = sparse(3*nC,9*nC);
    for j = 1:nC
        for k = 1:3
            idx_j = 3*(j-1); idx_jk = (k-1)*3*nC+3*(j-1);
            B(idx_j+1,idx_jk+2) = u(idx_jk+2);
            B(idx_j+2,idx_jk+3) = u(idx_jk+3);
            B(idx_j+3,idx_jk+2) = u(idx_jk+3);
            B(idx_j+3,idx_jk+3) = u(idx_jk+2);
        end
    end
    B(DNodes,:) = [];
    ZZZ = sparse(size(B,1),size(B,1));
    AAA = [(1+tau*alpha)*SSS_free,B(:,FNodes)';B(:,FNodes),ZZZ];
    rhs = -alpha*SSS*u+bbb;
    ddd = [rhs(FNodes);zeros(size(B,1),1)];
    xxx = AAA\ddd;
    dt_u(FNodes) = xxx(1:size(SSS_free,1));
    corr = sqrt(dt_u'*SSS*dt_u)
    u = u+tau*dt_u; show_p1_para(c4n,n4e,u);
end

function val = f3(x)
val = 0*ones(size(x,1),1);

function u = u_moebius(x)
L = max(x(:,1)); nX = size(x,1); u = zeros(9*nX,1);
u(0*nX+(1:3:3*nX)) = sin(2*pi*x(:,1)/L);
u(3*nX+(1:3:3*nX)) = x(:,2)+(1-2*x(:,2)).*sin(pi*x(:,1)/(2*L));
u(6*nX+(1:3:3*nX)) = sin(pi*x(:,1)/L);
u(0*nX+(2:3:3*nX)) = ones(nX,1);
u(3*nX+(3:3:3*nX)) = ones(nX,1)-2*(x(:,1)>L/2).*ones(nX,1);
```

Fig. 8.6 Approximation of the nonlinear Kirchhoff model with discrete Kirchhoff triangles

The matrix $D^\top S_2 D$ is generated as in the case of the linear model and provided by the routine `dkt_matrix.m`. The initial deformation is assumed to satisfy the boundary conditions which may be inhomogeneous. We refer to the implementation displayed in Fig. 8.6 for details.

8.4 Willmore Flow

We discuss in this section numerical methods for approximating the Willmore flow. This is the L^2-gradient flow of the Willmore energy which is defined on closed surfaces in $\mathbb{R}^3$. To compute the evolution equation, we review concepts from differential geometry to differentiate quantities on surfaces and to measure variations of surfaces. The reader is referred to the textbooks [13, 14] for further details. The numerical schemes are based on results in [2, 8, 9].

8.4.1 Tangential Differentiation and Curvature

Let $\mathscr{M} \subset \mathbb{R}^3$ be a surface, i.e., an orientable two-dimensional C^2-submanifold $\mathscr{M}$ in $\mathbb{R}^3$, with continuous unit normal $n : \mathscr{M} \to \mathbb{R}^3$. For scalar functions $f : \mathscr{M} \to \mathbb{R}$ and vector fields $F : \mathscr{M} \to \mathbb{R}^3$ on $\mathscr{M}$ that admit continuously differentiable extensions $\widetilde{f} : \mathscr{U}(\mathscr{M}) \to \mathbb{R}$ and $\widetilde{F} : \mathscr{U}(\mathscr{M}) \to \mathbb{R}^3$ to an open neighborhood of $\mathscr{M}$, we define the *tangential gradient* and the *tangential divergence* by

$$\nabla_{\mathscr{M}} f = \nabla \widetilde{f} - (n \cdot \nabla \widetilde{f}) n, \quad \operatorname{div}_{\mathscr{M}} F = \operatorname{div} \widetilde{F} - n^\top D\widetilde{F} n.$$

The operators satisfy the product rule

$$\operatorname{div}_{\mathscr{M}} (fF) = \nabla_{\mathscr{M}} f \cdot F + f \operatorname{div}_{\mathscr{M}} F.$$

The tangential gradient $\nabla_{\mathscr{M}} F$ of a vector field F is the matrix whose i-th row coincides with the transpose of the tangential gradient of the i-th component of F. The *Laplace–Beltrami operator* is defined as

$$\Delta_{\mathscr{M}} f = \operatorname{div}_{\mathscr{M}} \nabla_{\mathscr{M}} f.$$

For a local parametrization $u : \omega \to \mathbb{R}^3$ of $\mathscr{M}$, the tangent vectors $\partial_\ell u$, $\ell = 1, 2$, are linearly independent and define a unit normal $b = \pm \partial_1 u \times \partial_2 u / |\partial_1 u \times \partial_2 u|$, cf. Fig. 8.7. We assume in the following that the sign is chosen so that $b = n \circ u$. The *first fundamental form* is the matrix g with entries

$$g_{ij} = \partial_i u \cdot \partial_j u.$$

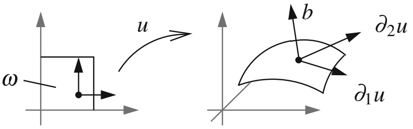


Fig. 8.7 Local parametrization of a surface by a mapping $u : \omega \to \mathbb{R}^3$; the partial derivatives $\partial_1 u$ and $\partial_2 u$ of u define a basis of the tangent space for every point on the image of u; their normalized cross product defines a unit normal b to the surface

It defines a metric on the tangent space of $\mathscr{M}$, e.g., the length of a tangent vector $\alpha_1\partial_1 u + \alpha_2\partial_2 u$ is given by the square root of $\alpha \cdot (g\alpha)$. The matrix g is symmetric and positive definite everywhere in ω; and we let $g^{-1} = (g^{ij})$ be its inverse and $g^{-1/2} = (g_{ij}^{(-1/2)})$ the symmetric and positive definite square root of g^{-1}.

Proposition 8.4 (Differential operators on $\mathscr{M}$) *We have*

$$(\nabla_{\mathscr{M}} f) \circ u = \sum_{i,j=1}^{2} g^{ij}\partial_j(f \circ u)\partial_i u, \quad (\operatorname{div}_{\mathscr{M}} F) \circ u = \sum_{i,j=1}^{2} g^{ij}\partial_j(F \circ u) \cdot \partial_i u.$$

If $F = \sum_{i=1}^{2} F_i \partial_i u$ is tangential or $F = \nabla_{\mathscr{M}} f$, then

$$(\operatorname{div}_{\mathscr{M}} F) \circ u = (\det g)^{-1/2} \sum_{i=1}^{2} \partial_i\big(F_i \circ u(\det g)^{1/2}\big),$$
$$(\Delta_{\mathscr{M}} f) \circ u = (\det g)^{-1/2} \sum_{i,j=1}^{2} \partial_i\big((\det g)^{1/2} g^{ij}\partial_j(f \circ u)\big).$$

In particular, the operators are independent of the extensions.

Proof We occasionally omit the composition with u, e.g., we write $\nabla_{\mathscr{M}} f$ for $(\nabla_{\mathscr{M}} f) \circ u$. For $k = 1, 2$ we have

$$(\nabla_{\mathscr{M}} f) \cdot \partial_k u = \nabla\widetilde{f} \cdot \partial_k u = \partial_k(\widetilde{f} \circ u) = \partial_k(f \circ u)$$

and $(\nabla_{\mathscr{M}} f) \cdot n = 0$. Since

$$\Big(\sum_{i,j=1}^{2} g^{ij}\partial_j(f \circ u)\partial_i u\Big) \cdot \partial_k u = \sum_{i,j=1}^{2} g^{ij} g_{ik}\partial_j(f \circ u) = \sum_{j=1}^{2} \delta_{jk}\partial_j(f \circ u) = \partial_k(f \circ u)$$

and since the sum on the right-hand side of the first asserted identity is orthogonal to n, we deduce the formula for $\nabla_{\mathscr{M}} f$. With $V_i = \sum_{j=1}^{2} g_{ij}^{(-1/2)}\partial_j u$ for $i = 1, 2$, the vectors (V_1, V_2, b) define an orthonormal basis in $\mathbb{R}^3$, i.e.,

$$V_i \cdot V_k = \sum_{j,\ell=1}^{2} g_{ij}^{(-1/2)} g_{k\ell}^{(-1/2)} \partial_j u \cdot \partial_\ell u = \sum_{j,\ell=1}^{2} g_{ij}^{(-1/2)} g_{k\ell}^{(-1/2)} g_{j\ell} = \delta_{ik}$$

and $V_i \cdot b = 0$ for $i = 1, 2$. With this we have

$$\operatorname{div} \widetilde{F} = \operatorname{tr} D\widetilde{F} = \sum_{i=1}^{2} V_i^\top D\widetilde{F} V_i + b^\top D\widetilde{F} b,$$

and hence by definition of $\operatorname{div}_{\mathscr{M}}$

$$\operatorname{div}_{\mathscr{M}} F = \sum_{i,j,k=1}^{2} g_{ij}^{(-1/2)} g_{ik}^{(-1/2)} (\partial_j u)^\top D\widetilde{F} \partial_k u = \sum_{j,k=1}^{2} g^{jk} \partial_j (F \circ u) \cdot \partial_k u$$

which is the second identity. Assume now that F is tangential so that $F \circ u = \sum_{i=1}^{2} F_i \partial_i u$ with uniquely defined functions $F_i : \omega \to \mathbb{R}$. It then follows that

$$\begin{aligned} \operatorname{div}_{\mathscr{M}} F &= \sum_{i,j,k=1}^{2} g^{ij} (\partial_j F_k \partial_k u + F_k \partial_j \partial_k u) \cdot \partial_i u \\ &= \sum_{i,j,k=1}^{2} g^{ij} (\partial_j F_k g_{ik} + F_k \partial_j \partial_k u \cdot \partial_i u) \\ &= \sum_{k=1}^{2} \big(\partial_k F_k + \sum_{i,j=1}^{2} g^{ij} F_k (\partial_k \partial_j u \cdot \partial_i u) \big). \end{aligned}$$

Since g^{-1} is symmetric, $g^{-1} = (\det g)^{-1} \det{}' g$, and $2\partial_k (\det g)^{1/2} = (\det g)^{-1/2} \det{}' g : \partial_k g$, we have for $k = 1, 2$ that

$$\sum_{i,j=1}^{2} g^{ij} (\partial_k \partial_j u \cdot \partial_i u) = \frac{1}{2} \sum_{i,j=1}^{2} g^{ij} \partial_k g_{ij} = (\det g)^{-1/2} \partial_k (\det g)^{1/2}.$$

The combination of the last two equations shows that

$$\operatorname{div}_{\mathscr{M}} F = \sum_{k=1}^{2} \big(\partial_k F_k + F_k (\det g)^{-1/2} \partial_k (\det g)^{1/2} \big),$$

which is the asserted identity. The identity for the Laplace–Beltrami operator now follows from the characterization of $\nabla_{\mathscr{M}}$. □

Example 8.1 For the parametrization $u(\theta,\phi) = r(\sin\theta\sin\phi, \sin\theta\cos\phi, \cos\theta)$ of the sphere $S_r \subset \mathbb{R}^3$ with radius $r > 0$, we have $\det g(\theta,\phi) = r^4 \sin^2\theta$ and $\Delta_{S_r} f = (r^2 \sin\theta)^{-1}\big[\partial_\theta(\sin\theta\,\partial_\theta f) + (\sin\theta)^{-1}\partial_\phi^2 f\big]$.

Remark 8.5 The representation $F = \sum_{i=1}^2 (V_i, F)V_i = \sum_{i,j=1}^2 g^{ij}(F \cdot \partial_i u)\partial_j u$ of a tangential vector field F with the orthonormal vectors (V_1, V_2) constructed in the proof of Proposition 8.4 yields the *Weingarten equation* $\partial_k b = -\sum_{i,j=1}^2 g^{ij} h_{ki} \partial_j u$ with the coefficients h_{ki} of the second fundamental form defined below.

To define a measure of curvature, we let $c : (-\varepsilon, \varepsilon) \to \mathscr{M}$ be a C^2 curve in $\mathscr{M}$ with $|c'(t)| = 1$ for all $t \in (-\varepsilon, \varepsilon)$ and consider the quantity $\kappa = c'' \cdot (n \circ c)$. Since $c' \cdot (n \circ c) = 0$ we have

$$\kappa = -c' \cdot (n \circ c)' = -c' \cdot \big(\nabla_{\mathscr{M}} n\, c'\big).$$

We call $\nabla_{\mathscr{M}} n$ the *shape operator* which is closely related to the *second fundamental form* defined through the symmetric matrix

$$h_{ij} = -\partial_i b \cdot \partial_j u = b \cdot \partial_i \partial_j u.$$

The mapping induced by $\nabla_{\mathscr{M}} n$ is also called the *Weingarten map*.

Proposition 8.5 (Shape operator) *The matrix $\nabla_{\mathscr{M}} n$ is symmetric and defines a self-adjoint linear operator on the tangent space of $\mathscr{M}$ into itself and is in the basis $(\partial_1 u, \partial_2 u)$ given by the generally nonsymmetric matrix $s = -hg^{-1}$.*

Proof For $i = 1, 2, 3$ we have $(\nabla_{\mathscr{M}} n_i) \cdot n = 0$ and hence $(\nabla_{\mathscr{M}} n)n = 0$. The identity $|n|^2 = 1$ implies that $n^\top(\nabla_{\mathscr{M}} n) = 0$. Therefore, $\nabla_{\mathscr{M}} n$ defines an endomorphism on the tangent space of $\mathscr{M}$; and for $i = 1, 2$ there exist s_{ij}, $j = 1, 2$, such that $(\nabla_{\mathscr{M}} n)\partial_i u = \sum_{j=1}^2 s_{ij} \partial_j u$, i.e.,

$$\sum_{j=1}^2 s_{ij} \partial_j u \cdot \partial_k u = (\nabla_{\mathscr{M}} n \partial_i u) \cdot \partial_k u = \partial_i (n \circ u) \cdot \partial_k u = \partial_i b \cdot \partial_k u = -h_{ik}$$

and hence with $\partial_j u \cdot \partial_k u = g_{jk}$ we deduce $sg = -h$. The identity also implies the symmetry of $\nabla_{\mathscr{M}} n$. □

The *principal curvatures* of $\mathscr{M}$ are the eigenvalues of the self-adjoint symmetric operator $\nabla_{\mathscr{M}} n$ restricted to the tangent space of $\mathscr{M}$ and are denoted by κ_1 and κ_2. The eigenvectors corresponding to κ_1 and κ_2 are called *directions of principal curvature*. The possibly nonsymmetric matrix s has the eigenvalues κ_1 and κ_2 and the *mean* and *Gauss curvature* are defined as

$$H = \operatorname{tr} s = \kappa_1 + \kappa_2, \quad K = \det s = \kappa_1 \kappa_2,$$

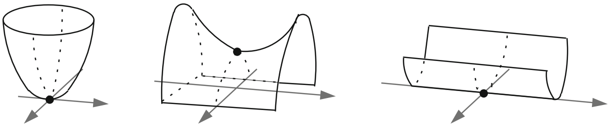


Fig. 8.8 Ellipsoidal surface with $\kappa_1 < 0$, $\kappa_2 < 0$ (*left*), hyperbolic surface with $\kappa_1 < 0$, $\kappa_2 > 0$ (middle), and parabolic surface with $\kappa_1 = 0$, $\kappa_2 > 0$ (*right*) relative to the unit normal $n = e_3$

respectively. We have that $|\nabla_{\mathscr{M}} n|^2 = s^\top : s = \operatorname{tr}(s^2) = \kappa_1^2 + \kappa_2^2 = (\operatorname{tr} s)^2 - 2 \det s = H^2 - 2K$. We also note the identities $H = -h : g^{-1} = \operatorname{tr}(-hg^{-1})$.

Remark 8.6 The sign of H depends on the choice of the unit normal, whereas K is independent of the sign of $\pm n$. The definition implies $\kappa_1, \kappa_2 \geq 0$ if $\mathscr{M}$ is locally convex with respect to the chosen unit normal. The mean curvature H is often defined as $(1/2) \operatorname{tr} s = (\kappa_1 + \kappa_2)/2$.

Typical local shapes of two-dimensional surfaces are given in the following example and are shown in Fig. 8.8.

Example 8.2 Consider a local parametrization of a surface that is given by the graph of the function $f : \omega \to \mathbb{R}$, i.e., $u(x) = \big(x, f(x)\big)$. Also assume that $0 \in \omega$ with $\nabla f(0) = 0$. Noting $\partial_i u = e_i$ for $i = 1, 2$, and $b = e_3$, $g = I$, and $h = b \cdot \partial_i \partial_j u = D^2 f$, we find that $s = -hg^{-1} = -D^2 f$ at $x = 0$.

Proposition 8.6 (Mean curvature) *We have*

$$\operatorname{div}_{\mathscr{M}} n = H, \quad -\Delta_{\mathscr{M}} \mathrm{id}_{\mathscr{M}} = Hn,$$

where $\mathrm{id}_{\mathscr{M}} : \mathscr{M} \to \mathbb{R}^3$ *denotes the identity on* $\mathscr{M}$, *i.e.,* $\mathrm{id}_{\mathscr{M}}(p) = p$ *for all* $p \in \mathscr{M}$ *and* $\Delta_{\mathscr{M}}$ *is applied to every component of* $\mathrm{id}_{\mathscr{M}}$.

Proof With the characterization of $\operatorname{div}_{\mathscr{M}}$ of Proposition 8.4, we have

$$\operatorname{div}_{\mathscr{M}} n = \sum_{i,j=1}^{m} g^{ij} \partial_j (n \circ u) \cdot \partial_i u = -\sum_{i,j=1}^{2} g^{ij} h_{ij} = -\operatorname{tr}(hg^{-1}) = \operatorname{tr} s.$$

We have $\nabla_{\mathscr{M}} \mathrm{id}_{\mathscr{M}} = I - nn^\top$ and thus $-\Delta_{\mathscr{M}} \mathrm{id}^i_{\mathscr{M}} = \operatorname{div}_{\mathscr{M}}(n^i n) = n^i H$. □

We have the following generalized integration-by-parts formula.

Proposition 8.7 (Integration-by-parts) *For a vector field* $F : \mathscr{M} \to \mathbb{R}^3$ *and a compactly supported function* $\varphi : \mathscr{M} \to \mathbb{R}$, *we have*

$$\int_{\mathscr{M}} \nabla_{\mathscr{M}} \varphi \cdot F \, \mathrm{d}s = -\int_{\mathscr{M}} \varphi \operatorname{div}_{\mathscr{M}} F \, \mathrm{d}s + \int_{\mathscr{M}} H(F \cdot n) \varphi \, \mathrm{d}s.$$

Proof We assume that φ belongs to a coordinate chart parametrized by u and consider the vector field $G = \varphi F$ on $\mathscr{M}$. We set $G = G_{\tan} + G_{\rm nor}$ with $G_{\rm nor} = \gamma n$ for $\gamma = G \cdot n$. Then $G_{\tan} = \sum_{i=1}^{2} G_i \partial_i u$ and Proposition 8.4 and an integration-by-parts in $\mathbb{R}^2$ yield

$$\int_{\mathscr{M}} \operatorname{div}_{\mathscr{M}} G_{\tan} \,\mathrm{d}s = \sum_{i=1}^{2} \int_{\omega} \partial_i (G_i (\det g)^{1/2}) \,\mathrm{d}x$$
$$= \int_{\omega} \operatorname{div} \big((\det g)^{1/2} [G_1, G_2]\big) \,\mathrm{d}x = 0.$$

The product rule and $(\nabla_{\mathscr{M}} \gamma) \cdot n = 0$ show that

$$\int_{\mathscr{M}} \operatorname{div}_{\mathscr{M}} G_{\rm nor} \,\mathrm{d}s = \int_{\mathscr{M}} \gamma \operatorname{div}_{\mathscr{M}} n \,\mathrm{d}s = \int_{\mathscr{M}} \gamma H \,\mathrm{d}s = \int_{\mathscr{M}} (G \cdot n) H \,\mathrm{d}s.$$

The combination of the identities and an application of the product rule prove the asserted formula. □

Remark 8.7 If φ does not vanish on the boundary of $\mathscr{M}$, then the boundary term $\int_{\partial \mathscr{M}} \varphi F \cdot \mu \,\mathrm{d}t$ with the conormal $\mu = \tau \times n$, where τ is the tangent on $\partial\omega$, has to be included on the right-hand side.

8.4.2 Normal Variations

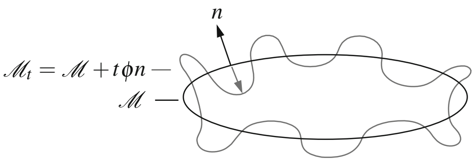
For a surface $\mathscr{M} \subset \mathbb{R}^3$ with unit normal n and a function $\phi : \mathscr{M} \to \mathbb{R}$, we consider for $-\varepsilon < t < \varepsilon$ the normal variations of $\mathscr{M}$ defined by

$$\mathscr{M}_t = \{q \in \mathbb{R}^3 : q = p + t\varphi(p)n(p),\ p \in \mathscr{M}\},$$

cf. Fig. 8.9. Then $\mathscr{M}_0 = \mathscr{M}$ and for sufficiently small $\varepsilon > 0$, the sets $\mathscr{M}_t$ are surfaces in $\mathbb{R}^3$. If $u : \omega \to \mathbb{R}^3$ is a local parametrization of $\mathscr{M}$, then

$$u_t = u + t(\phi \circ u)(n \circ u)$$

Fig. 8.9 Normal variation of a surface defined by a scalar function ϕ



is a local parametrization of $\mathscr{M}_t$. For a function $f_t : \mathscr{M}_t \to \mathbb{R}$ we denote $f = f_0$ and define

$$\delta f(p) = \lim_{t\to 0} t^{-1}\big(f_t(p) - f_0(p)\big)$$

for $p \in \mathscr{M}$. The proposition below studies the changes of geometric quantities on the surfaces $\mathscr{M}_t$ and employs Gauss' equation and an equivalent characterization of the Laplace–Beltrami operator stated in the following lemma.

Lemma 8.2 (Christoffel symbols) *With the* Christoffel symbols *of the first kind* $\Gamma_{ij,m} = \partial_i\partial_j u \cdot \partial_m u$ *and of the second kind* $\Gamma_{ij}^k = \sum_{m=1}^2 g^{km}\Gamma_{ij,m}$, *we have* Gauss' equation *and a representation of the Laplace–Beltrami operator, i.e.,*

$$\partial_i\partial_j u = \sum_{k=1}^2 \Gamma_{ij}^k \partial_k u + h_{ij} b, \quad \Delta_{\mathscr{M}}\phi = \sum_{i,j}^2 g^{ij}\Big(\partial_i\partial_j\phi - \sum_{k=1}^2 \Gamma_{ij}^k \partial_k\phi\Big).$$

Proof We have $\partial_i\partial_j u \cdot n = h_{ij}$ and hence there exist α_{ij}^k with

$$\partial_i\partial_j u \cdot \partial_\ell u = \sum_{k=1}^2 \alpha_{ij}^k \partial_k u \cdot \partial_\ell u = \sum_{k=1}^2 \alpha_{ij}^k g_{k\ell},$$

i.e., $\alpha_{ij}^m = \sum_{\ell=1}^2 g^{\ell m}(\partial_i\partial_j u)\cdot\partial_\ell u$. This implies the representation of $\partial_i\partial_j u$. According to Proposition 8.4 we have

$$\begin{aligned}
\Delta_{\mathscr{M}}\phi &= \sum_{i,j,\ell,m=1}^2 g^{ij}\partial_j\big(g^{\ell m}\partial_m\phi\partial_\ell u\big)\cdot\partial_i u \\
&= \sum_{i,j,\ell,m=1}^2 g^{ij}\big[\partial_j g^{\ell m}\partial_m\phi\partial_\ell u + g^{\ell m}(\partial_j\partial_m\phi)\partial_\ell u + g^{\ell m}\partial_m\phi(\partial_j\partial_\ell u)\big]\cdot\partial_i u \\
&= \sum_{i,j,\ell,m=1}^2 g^{ij}\big[\partial_j g^{\ell m}\partial_m\phi g_{\ell i} + g^{\ell m}(\partial_j\partial_m\phi)g_{\ell i} + g^{\ell m}\partial_m\phi\Gamma_{j\ell,i}\big].
\end{aligned}$$

Using $0 = \partial_j\sum_{r=1}^2(g^{\ell r}g_{rm}) = \sum_{r=1}^2(\partial_j g^{\ell r}g_{rm} + g^{\ell r}\partial_j g_{rm})$, we find that $\partial_j g^{\ell m} = -\sum_{r,k=1}^2 g^{\ell r}\partial_j g_{rk}g^{km}$ and noting $\partial_j g_{rk} = \Gamma_{jr,k} + \Gamma_{jk,r}$, i.e.,

$$\partial_j g^{\ell m} = -\sum_{r,k=1}^2 g^{\ell r}(\Gamma_{jr,k} + \Gamma_{jk,r})g^{km},$$

shows that $\Delta_{\mathscr{M}}\phi$ equals

$$\sum_{i,j,\ell,m=1}^{2} g^{ij}\big[-\sum_{r,k=1}^{2} g^{\ell r}(\Gamma_{jr,k}+\Gamma_{jk,r})g^{km}\partial_m\phi g_{\ell i}+g^{\ell m}(\partial_j\partial_m\phi)g_{\ell i}+g^{\ell m}\partial_m\phi\Gamma_{j\ell,i}\big]$$
$$=\sum_{i,j=1}^{2} g^{ij}\big[-\sum_{k,m=1}^{2}(\Gamma_{ji,k}+\Gamma_{jk,i})g^{km}\partial_m\phi+\partial_j\partial_i\phi+\sum_{\ell,m=1}^{2} g^{\ell m}\partial_m\phi\Gamma_{j\ell,i}\big]$$
$$=\sum_{i,j=1}^{2} g^{ij}\big[\partial_j\partial_i\phi-\sum_{k,m=1}^{2} g^{km}\Gamma_{ij,k}\partial_m\phi\big].$$

This implies the asserted formula for $\Delta_{\mathscr{M}}\phi$. □

A consequence of this is Gauss' *theorema egregium* which is stated below for isometric parametrizations, cf. Proposition 8.2.

Lemma 8.3 (Gauss curvature for isometries) *Assume that* $\Gamma_{ij,k}=\partial_i\partial_j u\cdot\partial_k u=0$ *for all* $1\le i,j,k\le 2$. *Then* $K=0$.

Proof Using $\partial_2(\partial_1^2 u)=\partial_1(\partial_1\partial_2 u)$ and the identities $\partial_i\partial_j u=h_{ij}b$, Lemma 8.2 shows that
$$0=\partial_2(h_{11}b)-\partial_1(h_{12}b)=(\partial_2 h_{11}-\partial_1 h_{12})b+h_{11}\partial_2 n-h_{12}\partial_1 n.$$

The Weingarten equations $\partial_k b=-\sum_{i,j=1}^2 g^{ij}h_{ki}\partial_j u$, cf. Remark 8.5, imply that for the tangential part of the identity, we have

$$0=-h_{11}\sum_{i,j=1}^{2} g^{ij}h_{2i}\partial_j u+h_{12}\sum_{i,j=1}^{2} g^{ij}h_{1i}\partial_j u=-\sum_{i,j=1}^{2} g^{ij}(h_{11}h_{2i}-h_{12}h_{1i})\partial_j u.$$

The contributions to the sum vanish for $i=1$ and hence

$$0=-(\det h)\sum_{j=1}^{2} g^{2j}\partial_j u.$$

Since $\partial_1 u$ and $\partial_2 u$ are linearly independent, this implies $\det h=0$ and $K=0$. □

Proposition 8.8 (Normal variations of geometric quantities) *For* $1\le i,j\le 2$ *we have*

$$\delta g_{ij}=-2\phi h_{ij},\quad \delta g_{ij}^{-1}=2\phi\sum_{k,\ell=1}^{2} g^{ik}h_{k\ell}g^{\ell j},\quad \delta(\det g)^{1/2}=\phi H(\det g)^{1/2}$$

and
$$\delta n=-\nabla_{\mathscr{M}}\phi,\quad \delta H=-\Delta_{\mathscr{M}}\phi-|s|^2.$$

Proof We identify ϕ with the function $\phi \circ u$ and write $b = n \circ u$. We also omit the dependence on t in the following. Noting $\partial_i b \cdot b = 0$, we have

$$g^t_{ij} = \partial_i u_t \cdot \partial_j u_t = g_{ij} + t\phi\big(\partial_i u \cdot \partial_j b + \partial_j u \cdot \partial_i b\big) + t^2 \partial_i \phi \partial_j \phi + t^2 \phi^2 \partial_i b \cdot \partial_j b,$$

which implies $\delta g_{ij} = -2\phi h_{ij}$. With $g^{-1} g = I_2$ we find that $\delta g^{-1} = -g^{-1}(\delta g) g^{-1}$ and hence

$$\delta g^{ij} = -\sum_{k,\ell=1}^{2} g^{ik}(\delta g_{k\ell}) g^{\ell j} = 2\phi \sum_{k,\ell=1}^{2} g^{ik} h_{k\ell} g^{\ell j}.$$

The relations $(\det g)^{-1} \det{}' g = g^{-1}$ and $g^{-1} : h = -H$ imply

$$\begin{aligned}\delta (\det g)^{1/2} &= \frac{1}{2}(\det g)^{-1/2}(\det{}' g) : \delta g = \frac{1}{2}(\det g)^{1/2} g^{-1} : \delta g \\ &= -\phi (\det g)^{1/2} g^{-1} : h = \phi (\det g)^{1/2} H.\end{aligned}$$

Using $b \cdot \partial_i u = 0$, we deduce $\delta b \cdot \partial_i u + b \cdot \delta \partial_i u = 0$ and with $\delta \partial_i u = \phi \partial_i b + (\partial_i \phi) b$ and $b \cdot \partial_i b = 0$, it follows that $\delta b \cdot \partial_i u = -\partial_i \phi$. Since $0 = \delta |b|^2 = 2\delta b \cdot b$, we have that there exist α_1, α_2 with $\delta b = \alpha_1 \partial_1 u + \alpha_2 \partial_2 u$. Noting

$$\sum_{i=1}^{2} \alpha_i \partial_i u \cdot \partial_k u = \delta b \cdot \partial_k u = -\partial_k \phi$$

we find that $\alpha_i = -\sum_{j=1}^{2} g^{ij} \partial_j \phi$ which implies

$$\delta b = -\sum_{i,j=1}^{2} g^{ij} \partial_j \phi \partial_i u,$$

and this expression coincides with $-\nabla_{\mathscr{M}} \phi$. It remains to compute δH. For this we first compute δh_{ij}. Noting

$$\delta \partial_i \partial_j u = (\partial_i \partial_j \phi) b + \partial_i \phi \partial_j b + \partial_j \phi \partial_i b + \phi \partial_i \partial_j b,$$

and using $b \cdot \partial_i \partial_j b = -\partial_i b \cdot \partial_j b$, we have

$$b \cdot (\delta \partial_i \partial_j u) = \partial_i \partial_j \phi - \phi \partial_i b \cdot \partial_j b.$$

The Weingarten equation $\partial_k b = \sum_{i,j=1}^{2} g^{ij} h_{ki} \partial_j u$ leads to

$$\partial_i b \cdot \partial_j b = \sum_{\ell,m,r,s=1}^{2} g^{\ell m} h_{i\ell} g^{rs} h_{jr} \partial_m u \cdot \partial_s u = \sum_{r,s=1}^{2} g^{rs} h_{is} h_{rj}.$$

The formula for δb and Gauss' equation show that

$$\delta b \cdot (\partial_i \partial_j u) = -\Big(\sum_{k,\ell=1}^{2} g^{k\ell} \partial_\ell \phi \partial_k u \Big) \cdot \Big(\sum_{m=1}^{2} \Gamma_{ij}^{m} \partial_m u \Big) = -\sum_{\ell=1}^{2} \Gamma_{ij}^{\ell} \partial_\ell \phi.$$

We thus have

$$\delta h_{ij} = (\delta b) \cdot \partial_i \partial_j u + b \cdot (\delta \partial_i \partial_j u) = -\sum_{\ell=1}^{2} \Gamma_{ij}^{\ell} \partial_\ell \phi + \partial_i \partial_j \phi - \phi \sum_{k,\ell=1}^{2} g^{k\ell} h_{i\ell} h_{kj}$$

and

$$\begin{aligned} \sum_{i,j=1}^{2} g^{ij} \delta h_{ij} &= \sum_{i,j=1}^{2} g^{ij} \Big(\partial_i \partial_j \phi - \sum_{\ell=1}^{2} \Gamma_{ij}^{\ell} \partial_\ell \phi \Big) - \phi \sum_{i,j,k,\ell=1}^{2} g^{ij} g^{k\ell} h_{i\ell} h_{kj} \\ &= \Delta_{\mathscr{M}} \phi - \phi |s|^2. \end{aligned}$$

For the mean curvature we find that

$$\begin{aligned} \delta H &= -\delta \sum_{i,j=1}^{2} g^{ij} h_{ij} \\ &= -\sum_{i,j=1}^{2} \big((\delta g^{ij}) h_{ij} + g^{ij} (\delta h_{ij}) \big) \\ &= -2\phi \sum_{i,j,k,\ell=1}^{2} g^{ik} h_{k\ell} g^{\ell j} h_{ij} - \Delta_{\mathscr{M}} \phi + \phi |s|^2 \\ &= -2\phi |s|^2 - \Delta_{\mathscr{M}} \phi + \phi |s|^2. \end{aligned}$$

This proves the proposition. □

We finally derive variations for functionals measuring the surface area and the enclosed volume by a surface. The variation of a functional $\mathscr{G}$ defined on C^2-surfaces is the limit

$$\delta \mathscr{G}(\mathscr{M})[\phi] = \lim_{t \to 0} t^{-1} \big(\mathscr{G}(\mathscr{M}_t) - \mathscr{G}(\mathscr{M}_0) \big)$$

for a surface $\mathscr{M}$ that is perturbed in the normal direction with a function ϕ as above.

Proposition 8.9 (Variations of area and volume functional) *For $\mathscr{M} = \partial \Omega$ define*

$$\mathscr{A}(\mathscr{M}) = \int_{\mathscr{M}} 1 \, \mathrm{d}s, \quad \mathscr{V}(\mathscr{M}) = \int_{\Omega} 1 \, \mathrm{d}\xi = \frac{1}{3} \int_{\mathscr{M}} s \cdot n \, \mathrm{d}s.$$

We have

$$\delta\mathscr{A}(\mathscr{M})[\phi] = \int_{\mathscr{M}} H\phi \,\mathrm{d}s, \quad \delta\mathscr{V}(\mathscr{M})[\phi] = \frac{1}{3}\int_{\mathscr{M}} (1+H)\phi \,\mathrm{d}s.$$

Proof The first identity is a direct consequence of Proposition 8.8. The second identity follows from $\mathrm{id}_{\mathscr{M}_t} \cdot n = t\phi$. □

8.4.3 Variation of the Willmore Functional

The normal variations of geometric quantities allow us to characterize stationary surfaces for the Willmore functional and to define related evolution problems. For a closed surface $\mathscr{M} \subset \mathbb{R}^3$, the bending energy is given by the *Willmore functional*

$$W(\mathscr{M}) = \frac{1}{2}\int_{\mathscr{M}} H^2 \,\mathrm{d}s.$$

The following theorem characterizes critical points of the functional.

Theorem 8.5 (Euler–Lagrange equations) *For a normal variation of $\mathscr{M}$ defined by a function $\phi : \mathscr{M} \to \mathbb{R}$, we have*

$$\delta W(\mathscr{M})[\phi] = \int_{\mathscr{M}} (-\Delta_{\mathscr{M}} H)\phi - |\nabla_{\mathscr{M}} n|^2 H\phi + \frac{1}{2}H^3\phi \,\mathrm{d}s,$$

where $|\nabla_{\mathscr{M}} n|^2 = H^2 - 2K$.

Proof We assume that ϕ is supported in a coordinate chart. We then have

$$\begin{aligned}\delta\frac{1}{2}\int_{\mathscr{M}} H^2 \,\mathrm{d}s &= \frac{1}{2}\delta\int_{\omega} H^2(\det g)^{1/2}\,\mathrm{d}x \\ &= \int_{\omega} H(\delta H)(\det g)^{1/2} + \frac{1}{2}H^2\delta(\det g)^{1/2}\,\mathrm{d}x \\ &= \int_{\omega} H(-\Delta_{\mathscr{M}}\phi - \phi|s|^2)(\det g)^{1/2} + \frac{1}{2}\phi H^3(\det g)^{1/2}\,\mathrm{d}x \\ &= \int_{\mathscr{M}} H(-\Delta_{\mathscr{M}}\phi) - \phi H|s|^2 + \frac{1}{2}\phi H^3\,\mathrm{d}s.\end{aligned}$$

Noting $|s|^2 = |\nabla_{\mathscr{M}} n|^2 = H^2 - 2K$ and integrating-by-parts proves the theorem. □

Definition 8.7 For a family of surfaces $(\mathscr{M}_t)_{t\in[0,T]}$ and a family of points on the surfaces given by a differentiable function $c : [0,T] \to \mathbb{R}^3$ with $c(t) \in \mathscr{M}_t$ for all $t \in [0,T]$ we define the *normal velocity* of $\mathscr{M}_t$ at $q_0 = c(t_0)$ by

$$V(q_0, t_0) = c'(t_0) \cdot n(q_0).$$

We let

$$(\phi, \psi)_{\mathscr{M}_t} = \int\limits_{\mathscr{M}_t} \phi\psi \, \mathrm{d}s$$

denote the L^2 inner product on $\mathscr{M}_t$.

Definition 8.8 (i) A family of surfaces $(\mathscr{M}_t)_{t\in[0,T]}$ evolves according to the *Willmore flow* if

$$(V(t), \phi)_{\mathscr{M}_t} = -\delta W(\mathscr{M}_t)[\phi]$$

for all $t \in [0,T]$ and all $\phi \in C^\infty(\mathscr{M}_t)$.
(ii) A family of surfaces $(\mathscr{M}_t)_{t\in[0,T]}$ evolves according to the *Helfrich flow* if there exist $\lambda, \mu : [0,T] \to \mathbb{R}$ such that

$$(V(t), \phi)_{\mathscr{M}_t} = -\delta W(\mathscr{M}_t)[\phi] + \lambda(t)\delta\mathscr{A}(\mathscr{M}_t)[\phi] + \mu(t)\delta\mathscr{V}(\mathscr{M}_t)[\phi]$$

for all $t \in [0,T]$ and all $\phi \in C^\infty(\mathscr{M}_t)$ and the mappings $t \mapsto \mathscr{A}(\mathscr{M}_t)$ and $t \mapsto \mathscr{V}(\mathscr{M}_t)$ are constant.

Remark 8.8 The existence of solutions for the Willmore and Helfrich flow is only understood in special situations, e.g., when the initial surface $\mathscr{M}_0$ is a small perturbation of a sphere.

8.4.4 Discretization of the Laplace–Beltrami Operator

For a surface $\mathscr{M} \subset \mathbb{R}^3$, let $\mathscr{M}_h$ be an approximate surface that is the union of flat triangles in the triangulation $\mathscr{T}_h$ with vertices $\mathscr{N}_h \subset \mathbb{R}^3$, cf. Fig. 8.10. The elementwise constant unit normal n_h on $\mathscr{M}_h$ defines the tangential gradient of a function $v_h \in \mathscr{S}^1(\mathscr{T}_h)$ via

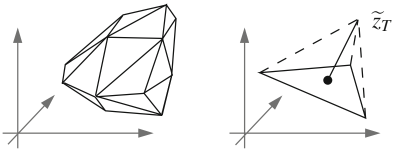


Fig. 8.10 Triangulated surface (*left*) and construction of an auxiliary tetrahedron with the auxiliary node $\widetilde{z}_T = x_T + |T|^{1/2} n_T$ (*right*)

$$\nabla_{\mathscr{M}_h} v_h = P_h \nabla \widetilde{v}_h = \big(I - n_h \otimes n_h\big) \nabla \widetilde{v}_h,$$

where $\widetilde{v}_h$ is an arbitrary extension of v_h to $\mathbb{R}^3$, e.g., by introducing for each triangle $T \in \mathscr{T}_h$ the auxiliary node $\widetilde{z}_T = x_T + |T|^{1/2} n_h|_T$, cf. Fig. 8.10, and setting $\widetilde{v}_h(\widetilde{z}_T) = 0$. The Laplace–Beltrami operator on a surface $\mathscr{M}$ leads to a Poisson problem on $\mathscr{M}$ of the form

$$-\Delta_{\mathscr{M}} u = f \text{ on } \mathscr{M}, \quad u = u_{\mathrm{D}} \text{ on } \gamma_{\mathrm{D},h}, \quad \nabla_{\mathscr{M}_h} u \cdot \mu_h = g \text{ on } \gamma_{\mathrm{N},h},$$

where μ_h is the conormal on $\Gamma_{\mathrm{N},h} \subset \partial \mathscr{M}_h$. A discrete approximation seeks $u_h \in \mathscr{S}^1(\mathscr{T}_h)$ such that $u_h|_{\gamma_{\mathrm{D},h}} = u_{\mathrm{D},h}$

$$\int_{\mathscr{M}_h} \nabla_{\mathscr{M}_h} u_h \cdot \nabla_{\mathscr{M}_h} v_h \,\mathrm{d}s = \int_{\mathscr{M}_h} f v_h \,\mathrm{d}s + \int_{\gamma_{\mathrm{N},h}} g_h v_h \,\mathrm{d}t$$

for all $v_h \in \mathscr{S}^1(\mathscr{T}_h)$ with $v_h|_{\gamma_{\mathrm{D},h}} = 0$. If $\gamma_{\mathrm{D},h} = \emptyset$, then the condition $\int_{\mathscr{M}_h} u_h \,\mathrm{d}s = 0$ is imposed. The MATLAB code displayed in Fig. 8.11 realizes the numerical scheme for the Laplace–Beltrami operator.

8.4.5 A Numerical Scheme for the Willmore Flow

We recall that the Willmore flow for a given initial surface $\mathscr{M}_0 \subset \mathbb{R}^3$ seeks a family of surfaces $(\mathscr{M}_t)_{t \in [0,T]}$ that solve the equation

$$V = \Delta_{\mathscr{M}} H + H |\nabla_{\mathscr{M}} n|^2 - \frac{1}{2} H^3,$$

where V is the normal velocity of $(\mathscr{M}_t)_{t \in [0,T]}$, n a unit normal on $\mathscr{M}_t$, and H the mean curvature of $\mathscr{M}_t$. For the position vector $X : \mathscr{M}_t \to \mathbb{R}^3$ on $\mathscr{M}$, we have $V = (\partial_t X) \cdot n$ and $Hn = -\Delta_{\mathscr{M}} \mathrm{id}_{\mathscr{M}}$. To discretize the evolution equation we consider a time step $t_k \in [0, T]$ and assume that we are given a triangulation $\mathscr{T}_h^k$ that defines the closed polyhedral surface $\mathscr{M}_h^k$ with unit normal $n_h^k \in \mathscr{L}^0(\mathscr{T}_h)^3$. We also suppose that $\widetilde{n}_h^k \in \mathscr{S}^1(\mathscr{T}_h^k)^3$ and $H_h^k \in \mathscr{S}^1(\mathscr{T}_h^k)$ approximate the unit normal n and the mean curvature of a smooth approximation of $\mathscr{M}_h^k$. To define the new surface $\mathscr{M}_h^{k+1}$, we compute a mapping

$$X_h^{k+1} : \mathscr{M}_h^k \to \mathbb{R}^3$$

```
function laplace_beltrami(red)
[c4n,n4e,Db,Nb] = triang_torus(.5,1,red);
nE = size(n4e,1); nC = size(c4n,1);
nNb = size(Nb,1); nDb = size(Db,1);
dNodes = unique(Db); fNodes = setdiff(1:nC,dNodes);
max_ctr = 9*nE; ctr = 0;
I = zeros(max_ctr,1); J = zeros(max_ctr,1);
X_s = zeros(max_ctr,1);
b = zeros(nC,1); c = zeros(nC,1); u = zeros(nC,1);
for j = 1:nE
    n_T = cross(c4n(n4e(j,2),:)-c4n(n4e(j,1),:),...
        c4n(n4e(j,3),:)-c4n(n4e(j,2),:));
    area_T = norm(n_T)/2;
    n_T = n_T/norm(n_T);
    mp_T = sum(c4n(n4e(j,:),:))/3;
    aux_tetra = [c4n(n4e(j,:),:);mp_T+sqrt(area_T)*n_T];
    grads3_T = [1,1,1,1;aux_tetra']\[0,0,0;eye(3)];
    P_T = eye(3)-n_T'*n_T;
    for k = 1:3
        b(n4e(j,k)) = b(n4e(j,k))+(1/3)*area_T*f(mp_T);
        c(n4e(j,k)) = c(n4e(j,k))+(1/3)*area_T;
        for ell = 1:3
            ctr = ctr+1;
            I(ctr) = n4e(j,k); J(ctr) = n4e(j,ell);
            X_s(ctr) = area_T*(P_T*grads3_T(k,:)')'...
                *(P_T*grads3_T(ell,:)');
        end
    end
end
s = sparse(I,J,X_s,nC,nC);
for j = 1:nNb
    length_E = norm(c4n(Nb(j,1),:)-c4n(Nb(j,2),:));
    mp_E = (c4n(Nb(j,1),:)-c4n(Nb(j,2),:))/2;
    b(Nb(j,1)) = b(Nb(j,1))+(1/2)*length_E*g(mp_E);
    b(Nb(j,2)) = b(Nb(j,2))+(1/2)*length_E*g(mp_E);
end
if isempty(dNodes)
    s = [s,c;c',0]; b = [b;0];
else
    for j = 1:nDb
        u(dNodes(j)) = u_D(c4n(dNodes(j),:));
    end
    b = b-s*u;
end
u(fNodes) = s(fNodes,fNodes)\b(fNodes);
show_p1_surf(c4n,n4e,u);

function val = f(X); val = X(2);
function val = u_D(X); val = 0;
function val = g(X); val = 0;
```

Fig. 8.11 MATLAB routine for the approximation of the Poisson problem on a surface

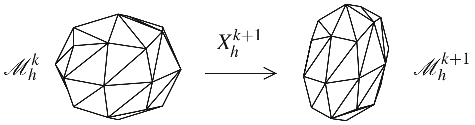


Fig. 8.12 Deformation $X_h^{k+1} : \mathscr{M}_h^k \to \mathbb{R}^3$ of a surface $\mathscr{M}_h^k$ that defines the new surface $\mathscr{M}_h^{k+1}$

that defines $\mathscr{M}_h^{k+1} = X_h^{k+1}(\mathscr{M}_h^k)$, cf. Fig. 8.12. A function or vector field on $\mathscr{M}_h^k$ is identified with a function on $\mathscr{M}_h^{k+1}$ via the parametrization X_h^{k+1}. The vector field $X_h^{k+1} \in \mathscr{S}^1(\mathscr{T}_h^k)^3$ is obtained by the following semi-implicit discretization of the Willmore flow from [2].

Algorithm 8.2 (Discrete Willmore flow) *For a discrete surface $\mathscr{M}_h^0$, functions $\widetilde{n}_h^0 \in \mathscr{S}^1(\mathscr{T}_h^0)^3$ and $H_h^0 = \mathscr{A}_h^0 \operatorname{div}_{\mathscr{M}_h^0} \widetilde{n}_h^0$. and a step size $\tau > 0$, compute the sequence $(\mathscr{M}_h^k)_{k=0,\dots,K}$ via $\mathscr{M}_h^{k+1} = X_h^{k+1}(\mathscr{M}_h^k)$, where $X_h^{k+1} \in \mathscr{S}^1(\mathscr{T}_h^k)^3$ and $H_h^{k+1} \in \mathscr{S}^1(\mathscr{T}_h^k)$ solve*

$$\frac{1}{\tau}\big(X_h^{k+1} - \mathrm{id}_{\mathscr{M}_h^k}, v_h \widetilde{n}_h^k\big)_{k,h} + \big(\nabla_{\mathscr{M}_h^k} H_h^{k+1}, \nabla_{\mathscr{M}_h^k} v_h\big)_k + \frac{1}{2}\big(|H_h^k|^2 H_h^{k+1}, v_h\big)_{k,h}$$
$$= \big(H_h^k \mathscr{A}_h^k |\nabla_{\mathscr{M}_h^k} \widetilde{n}_h^k|^2, v_h\big)_{k,h},$$
$$\big(H_h^{k+1} \widetilde{n}_h^k, Y_h\big)_{k,h} - \big(\nabla_{\mathscr{M}_h^k} X_h^{k+1}, \nabla_{\mathscr{M}_h^k} Y_h\big)_k = 0$$

for all $v_h \in \mathscr{S}^1(\mathscr{T}_h^k)$ and $Y_h \in \mathscr{S}^1(\mathscr{T}_h^k)^3$, and set $\widetilde{n}_h^{k+1} = \mathscr{A}_h^{k+1} n_h^{k+1}$. Stop the iteration if $\|v_h^{k+1}\|_{h,k} \le \varepsilon_{\mathrm{stop}}$ for $V_h^{k+1} = (X_h^{k+1} - \mathrm{id}_{\mathscr{M}_h^k})/\tau$ and $v_h^{k+1} = V_h^{k+1} \cdot \widetilde{n}_h^k$.

The averaging operator $\mathscr{A}_h^k : L^1(\mathscr{M}_h^k) \to \mathscr{S}^1(\mathscr{T}_h^k)$ is defined through

$$\mathscr{A}_h^k v(z) = \frac{1}{|\omega_z|} \sum_{T \in \mathscr{T}_h^k,\, z \in T} |T|\, v|_T, \qquad |\omega_z| = \sum_{T \in \mathscr{T}_h^k,\, z \in T} |T|,$$

and the inner product $(\cdot, \cdot)_{k,h}$ is for $v, w \in C(\mathscr{M}_h^k)$ defined by

$$(v, w)_{k,h} = \int_{\mathscr{M}_h^k} \mathscr{I}_h^k[vw]\, \mathrm{d}x.$$

Remark 8.9 The precise stability and convergence properties of Algorithm 8.2 are not known. The algorithm has an equidistribution property in the sense that it equidistributes the nodes of the discrete surface which avoids mesh irregularities. Details are discussed in [2].

According to Proposition 8.9 it suffices to impose that

$$\int_{\mathscr{M}} V \,\mathrm{d}s = \int_{\mathscr{M}} VH \,\mathrm{d}s = 0$$

to guarantee that the surface area and the enclosed volume are preserved. This leads to an identity for the associated Lagrange multipliers in the evolution equation, i.e.,

$$V = \Delta_{\mathscr{M}} H + H|\nabla_{\mathscr{M}} n|^2 - \frac{1}{2}H^3 + \lambda H + \mu.$$

Testing the equation with a constant function and with $H - \overline{H}$, where $\overline{H}$ is the integral mean of H, leads to

$$\mu = \frac{1}{|\mathscr{M}|} \int_{\mathscr{M}} -H|\nabla_{\mathscr{M}} n|^2 + \frac{1}{2}H^3 - \lambda H \,\mathrm{d}s,$$

$$\lambda = \frac{\int_{\mathscr{M}} \big(- H|\nabla_{\mathscr{M}} n|^2 + \frac{1}{2}H^3\big)(H - \overline{H}) + |\nabla_{\mathscr{M}} H|^2 \,\mathrm{d}s}{\int_{\mathscr{M}} (H - \overline{H})^2 \,\mathrm{d}s}.$$

To incorporate the constraints in Algorithm 8.2, the term λH is discretized implicitly if $\lambda \geq 0$ and explicitly otherwise. The MATLAB implementation displayed in Fig. 8.14 requires the bilinear forms

$$(\varphi_z^\ell, \varphi_y)_{k,h}, \quad (\nabla \varphi_z, \nabla \varphi_y)_k, \quad (\nabla \varphi_z^\ell, \nabla \varphi_y^m)_k,$$
$$(\varphi_z^\ell, n\varphi_y)_{k,h}, \quad (\varphi_z \mathscr{A}_h^k |\nabla_{\mathscr{M}_h^k} \widetilde{n}_h^k|^2, \varphi_y)_{k,h}, \quad (|H_h^k|^2 \varphi_z, \varphi_y)_{k,h},$$

for pairs of nodes $z, y \in \mathscr{N}_h^k$ and associated scalar nodal basis functions $\varphi_z, \varphi_y \in \mathscr{S}^1(\mathscr{T}_h)^k$ and vectorial nodal basis functions $\varphi_z^\ell = \varphi_z e_\ell$ and $\varphi_y^m = \varphi_y e_m$ with the canonical basis vectors $e_\ell, e_m \in \mathbb{R}^3$. The representing matrices are encoded in the arrays `m`, `s`, `S`, `M_n`, `m_w` provided by the routine shown in Fig. 8.13 while the last one is directly computed and stored in the array `m_H`. The routine `willmore_matrices.m` also computes an approximation of the mean curvature through $H_h^k = \mathscr{A}_h^k(\operatorname{div}_{\mathscr{M}_h^k} \widetilde{n}_h^k)$.

```
function [m,s,S,M_n,m_w,H] = willmore_matrices(c4n,n4e,w)
nC = size(c4n,1); nE = size(n4e,1);
max_ctr = 9*nE; ctr = 0;
I = zeros(max_ctr,1); J = zeros(max_ctr,1);
X_s = zeros(max_ctr,1);
diag_m = zeros(nC,1);
diag_m_w = zeros(nC,1);
diag_M_n = zeros(nC,3);
tr_nabla_w = zeros(nC,1);
for j = 1:nE
    n_T = cross(c4n(n4e(j,2),:)-c4n(n4e(j,1),:),...
        c4n(n4e(j,3),:)-c4n(n4e(j,2),:));
    area_T = norm(n_T)/2;
    n_T = n_T/norm(n_T);
    mp_T = sum(c4n(n4e(j,:),:))/3;
    tmp_tetra = [c4n(n4e(j,:),:);mp_T+sqrt(area_T)*n_T];
    grads3_T = [1,1,1,1;tmp_tetra']\[0,0,0;eye(3)];
    P_T = eye(3)-n_T'*n_T;
    P_Dphi_T = grads3_T(1:3,:)*P_T;
    nabla_T_w = w(n4e(j,:),:)'*P_Dphi_T;
    tr_nabla_w(j) = trace(nabla_T_w);
    W_sq = sum(sum(nabla_T_w.^2));
    for k = 1:3
        diag_m(n4e(j,k)) = diag_m(n4e(j,k))+area_T/3;
        diag_m_w(n4e(j,k)) = diag_m_w(n4e(j,k))+area_T*W_sq/3;
        diag_M_n(n4e(j,k),:) = diag_M_n(n4e(j,k),:)...
            +(area_T/3)*n_T;
        for ell = 1:3
            ctr = ctr+1;
            I(ctr) = n4e(j,k); J(ctr) = n4e(j,ell);
            X_s(ctr) = area_T...
                *(P_T*grads3_T(k,:)')'*(P_T*grads3_T(ell,:)');
        end
    end
end
m = spdiags(diag_m,0,nC,nC); m_w = spdiags(diag_m_w,0,nC,nC);
II = [3*I-2;3*I-1;3*I]; JJ = [3*J-2;3*J-1;3*J];
s = sparse(I,J,X_s); S = sparse(II,JJ,repmat(X_s,3,1));
I = [1:3:3*nC,2:3:3*nC,3:3:3*nC]'; J = [1:nC,1:nC,1:nC]';
M_n = sparse(I,J,diag_M_n(:));
H = average_quant_surf(c4n,n4e,tr_nabla_w);
```

Fig. 8.13 Matrices required in the implementation of the Willmore and the Helfrich flow

```
function willmore_helfrich_flow(red)
[n4e,c4n,¬,¬] = triang_sphere(red);
c4n(:,3) = .4*c4n(:,3);
tau = 2^(-red)/200;
nC = size(c4n,1);
w = averaged_normal(c4n,n4e);
[¬,¬,¬,¬,¬,H] = willmore_matrices(c4n,n4e,w);
X = reshape(c4n',3*nC,1);
corr = 1; eps_stop = 1e-1;
while corr > eps_stop
    w = averaged_normal(c4n,n4e);
    [m,s,S,M_n,m_w,¬] = willmore_matrices(c4n,n4e,w);
    m_H = spdiags(diag(m).*H.^2,0,nC,nC);
    [lambda,mu] = helfrich_constraints(c4n,H,s,m,m_w,m_H);
    A = [M_n',tau*(s+m_H/2-max(lambda,0)*m);-S,M_n];
    b = [tau*m_w*H+M_n'*X+tau*(mu*m*ones(nC,1)...
        +min(lambda,0)*m*H);zeros(3*nC,1)];
    xx = A\b;
    V = (xx(1:3*nC)-X)/tau;
    v = sum(reshape(V',3,nC)'.*w,2);
    corr = sqrt(v'*m*v)
    H = xx(3*nC+(1:nC)); X = X+tau*V; c4n = reshape(X',3,nC)';
    show_p1_surf(c4n,n4e,H);
end

function [lambda,mu] = helfrich_constraints(c4n,H,s,m,m_w,m_H)
nC = size(c4n,1); I = ones(nC,1);
mean_H = I'*m*H/(I'*m*I);
q = (H-mean_H)'*m*(H-mean_H);
lambda = 0;
if q > 0
    lambda = (-H'*m_w*H+H'*m_H/2*H...
        -(-H'*m_w*I+H'*m_H/2*I)*mean_H+H'*s*H)/q;
end
mu = (-H'*m_w*I+I'*m_H/2*H-lambda*I'*m*H)/(I'*m*I);

function w = averaged_normal(c4n,n4e)
nC = size(c4n,1); nE = size(n4e,1);
n = zeros(nE,3); w = zeros(nC,3);
for j = 1:nE
    n_T = cross(c4n(n4e(j,2),:)-c4n(n4e(j,1),:),...
        c4n(n4e(j,3),:)-c4n(n4e(j,2),:));
    n(j,:) = n_T/norm(n_T);
end
for k = 1:3
    w(:,k) = average_quant_surf(c4n,n4e,n(:,k));
end
norm_w = sqrt(sum(w.^2,2));
w = w./(norm_w*ones(1,3));
```

Fig. 8.14 Numerical approximation of the Willmore and the Helfrich flow

References

1. Arnold, D.N., Falk, R.S.: A uniformly accurate finite element method for the Reissner-Mindlin plate. SIAM J. Numer. Anal. **26**(6), 1276–1290 (1989). http://dx.doi.org/10.1137/0726074
2. Barrett, J.W., Garcke, H., Nürnberg, R.: Parametric approximation of Willmore flow and related geometric evolution equations. SIAM J. Sci. Comput. **31**(1), 225–253 (2008). http://dx.doi.org/10.1137/070700231
3. Bartels, S.: Approximation of large bending isometries with discrete Kirchhoff triangles. SIAM J. Numer. Anal. **51**(1), 516–525 (2013). http://dx.doi.org/10.1137/110855405
4. Boffi, D., Brezzi, F., Fortin, M.: Mixed Finite Element Methods and Applications. Springer Series in Computational Mathematics, vol. 44. Springer, Heidelberg (2013)
5. Braess, D.: Finite Elements, 3rd edn. Cambridge University Press, Cambridge (2007)
6. Ciarlet, P.G.: Mathematical Elasticity. Vol. II: Theory of Plates, Studies in Mathematics and Its Applications, vol. 27. North-Holland Publishing, Amsterdam (1997)
7. Conti, S.: Derivation of nonlinear plate models (2009). personal communication
8. Dziuk, G.: Finite elements for the Beltrami operator on arbitrary surfaces. In: Partial Differential Equations and Calculus of Variations. Lecture Notes in Math., vol. 1357, pp. 142–155. Springer, Berlin (1988). http://dx.doi.org/10.1007/BFb0082865
9. Dziuk, G.: Computational parametric Willmore flow. Numer. Math. **111**(1), 55–80 (2008). http://dx.doi.org/10.1007/s00211-008-0179-1
10. Friesecke, G., James, R.D., Müller, S.: A theorem on geometric rigidity and the derivation of nonlinear plate theory from three-dimensional elasticity. Commun. Pure Appl. Math. **55**(11), 1461–1506 (2002). http://dx.doi.org/10.1002/cpa.10048
11. Girault, V., Raviart, P.A.: Finite Element Methods for Navier-Stokes Equations, Springer Series in Computational Mathematics, vol. 5. Springer, Berlin (1986)
12. Hornung, P.: Approximating $W^{2,2}$ isometric immersions. C. R. Math. Acad. Sci. Paris **346**(3–4), 189–192 (2008). http://dx.doi.org/10.1016/j.crma.2008.01.001
13. Kühnel, W.: Differential Geometry. Student Mathematical Library, vol. 16. American Mathematical Society, Providence (2002)
14. Willmore, T.J.: Riemannian geometry. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York (1993)

Part III
Methods for Extended Formulations

Chapter 9
Nonconvexity and Microstructure

9.1 Analytical Properties

We discuss in this section features of minimization problems for energy functionals of the form

$$I(u) = \int_{\Omega} W(\nabla u)\,\mathrm{d}x - \int_{\Omega} f \cdot u\,\mathrm{d}x - \int_{\Gamma_{\mathrm{N}}} g \cdot u\,\mathrm{d}s$$

with a continuous but nonconvex energy density $W : \mathbb{R}^{m \times d} \to \mathbb{R}$ that is assumed to be nonnegative and to satisfy a p-growth condition for some $p > 1$. Although the functional I is coercive and bounded from below, the direct method in the calculus of variations cannot be applied due to the lack of weak lower semicontinuity of I for which convexity or quasiconvexity is required. In fact, infimizing sequences that are bounded exist but the energy functional may have no minimizers. Two natural questions arise:

- Do the weak limits of infimizing sequences solve a well-posed modified problem and can these be approximated numerically?
- Do the infimizing sequences contain information that explain the failure of weak lower semicontinuity and are these accessible?

It turns out that the weak limits of infimizing sequences are exactly the minimizers of the functional I^{qc} that is obtained from I by replacing W by its quasiconvex envelope. Since I is strongly continuous, the failure of weak lower semicontinuity is precisely related to the occurrence of oscillations that prohibit strong convergence. These oscillations are physically meaningful and of importance in applications. They can be efficiently described in a statistical sense with the help of measure-valued mappings. The ill-posed minimization of I may result from neglecting a higher order term in a well-posed minimization problem, e.g., in

$$I_{\varepsilon}(u) = \frac{\varepsilon}{2}\|D^2 u\|^2 + \int_{\Omega} W(\nabla u)\,\mathrm{d}x - \int_{\Omega} f \cdot u\,\mathrm{d}x - \int_{\Omega} g \cdot u\,\mathrm{d}s$$

S. Bartels, *Numerical Methods for Nonlinear Partial Differential Equations*,
Springer Series in Computational Mathematics 47,
DOI 10.1007/978-3-319-13797-1_9

with a small parameter $\varepsilon > 0$. The motivation for this is that the scale introduced by ε is too small to be resolved by numerical solution methods. Due to the nonconvexity of W, the gradient ∇u oscillates between different values, describing certain microstructures. The quadratic term involving the Hessian of u controls the frequency of these oscillations. When this term is neglected the oscillations become arbitrarily fast leading to infimizing sequences that are not strongly convergent. We discuss these effects in simplified model situations and refer the reader to the textbooks [8, 15], the survey articles [13, 14], and the seminal paper [2] for further details.

9.1.1 A Scalar Model Problem

Most of the problems related to nonconvex energy minimization become apparent for scalar and even one-dimensional problems. For $\Omega \subset \mathbb{R}^d$ we first consider the functional

$$I(u) = \int_\Omega W(\nabla u)\,\mathrm{d}x$$

with the energy density $W : \mathbb{R}^d \to \mathbb{R}$ defined for $F \in \mathbb{R}^d$ by

$$W(F) = \frac{1}{4}(|F|^2 - 1)^2,$$

the set of admissible functions

$$\mathscr{A} = \{u \in W^{1,4}(\Omega) : u|_{\Gamma_{\mathrm{D}}} = u_{\mathrm{D}}\}$$

for a possibly empty set $\Gamma_{\mathrm{D}} \subset \partial\Omega$, and a function $u_{\mathrm{D}} = \widetilde{u}_{\mathrm{D}}|_{\Gamma_{\mathrm{D}}}$ for some $\widetilde{u}_{\mathrm{D}} \in W^{1,4}(\Omega)$. We note that the convex hull W^{**} of W is for $F \in \mathbb{R}^d$ given by

$$W^{**}(F) = \begin{cases} W(F) & \text{for} \quad |F| \ge 1, \\ 0 & \text{for} \quad |F| \le 1. \end{cases}$$

The convex hull W^{**} is the largest convex function below W and is obtained by computing twice the Fenchel conjugate of W, cf. Fig. 9.1.

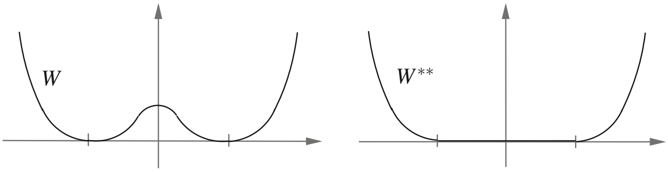


Fig. 9.1 Function $W(F) = (|F|^2 - 1)^2/4$ and its convex hull $W^{**}(F)$ that coincides with W for $|F| \ge 1$ and vanishes otherwise

The following proposition discusses the existence of solutions and infimizing sequences for affine boundary conditions. These play a special role in the relaxation of nonconvex minimization problems.

Proposition 9.1 (Affine boundary conditions) *For $\Gamma_{\mathrm{D}} = \partial\Omega$ and the affine boundary condition $\widetilde{u}_{\mathrm{D}}(x) = \overline{F} \cdot x$, for $\overline{F} \in \mathbb{R}^d$ and $x \in \Omega$, the functional I has the unique minimizer $u = \widetilde{u}_{\mathrm{D}} \in \mathscr{A}$ satisfying $I(u) = |\Omega| W^{**}(\overline{F}) = |\Omega| W(\overline{F})$ if $|\overline{F}| \geq 1$. If $|F| < 1$, we have $\inf_{u \in \mathscr{A}} I(u) = |\Omega| W^{**}(\overline{F}) < |\Omega| W(\overline{F})$ and there exists a bounded infimizing sequence $(u_j)_{j\in\mathbb{N}} \subset \mathscr{A} \cap W^{1,\infty}(\Omega)$ with $\|u_j - \widetilde{u}_{\mathrm{D}}\|_{L^\infty(\Omega)} \to 0$.*

Proof The proof follows from the observation that W is convex only in $\mathbb{R}^d \setminus B_1(0)$ in the sense that

$$DW(F) \cdot (G - F) + W(F) \leq W(G)$$

holds for all $G \in \mathbb{R}^d$ if and only if $|F| \geq 1$. If $|\overline{F}| \geq 1$, then due to the above inequality the function $u(x) = \overline{F} \cdot x$ satisfies for every $v \in \mathscr{A}$

$$\int_\Omega W(\nabla u)\,\mathrm{d}x \leq \int_\Omega W(\nabla v)\,\mathrm{d}x + \int_\Omega DW(\nabla u) \cdot \nabla(u - v)\,\mathrm{d}x.$$

Since $\nabla u = \overline{F}$ is constant and $(u - v)|_{\partial\Omega} = 0$, we have

$$\int_\Omega DW(\nabla u) \cdot \nabla(u - v)\,\mathrm{d}x = 0.$$

Therefore, u is a minimizer with $I(u) = |\Omega| W(\overline{F}) = |\Omega| W^{**}(\overline{F})$. If $|\overline{F}| < 1$, we claim that there exists a sequence $(u_j)_{j\in\mathbb{N}} \subset \mathscr{A}$ such that $I(u_j) \to 0$ as $j \to \infty$. To construct the sequence $(u_j)_{j\in\mathbb{N}}$, we note that there exist $F_1, F_2 \in \mathbb{R}^d$ and $\theta \in (0, 1)$ with $|F_1| = |F_2| = 1$ and $\overline{F} = \theta F_1 + (1 - \theta) F_2$. For $j \in \mathbb{N}$ we define $\widetilde{u}_j \in W^{1,\infty}(\mathbb{R}^d)$ by

$$\widetilde{u}_j(x) = F_1 \cdot x + \int_0^{(F_2 - F_1)\cdot x} \widetilde{\chi}_{(\theta,1)}(js)\,\mathrm{d}s,$$

where $\widetilde{\chi}_{(\theta,1)} : \mathbb{R} \to \mathbb{R}$ is the one-periodic extension of the characteristic function $\chi_{(\theta,1)} : (0, 1) \to \{0, 1\}$ of the interval $(\theta, 1)$. Figure 9.2 illustrates the construction.

For every $j \geq 1$ the function $\widetilde{u}_j$ satisfies

$$\begin{aligned}\nabla\widetilde{u}_j(x) &= F_1 + \widetilde{\chi}_{(\theta,1)}\big(j(F_2 - F_1)\cdot x\big)(F_2 - F_1)\\ &= \begin{cases} F_1 & \text{if} \quad k \leq j(F_2 - F_1)\cdot x \leq k + \theta,\ k \in \mathbb{Z},\\ F_2 & \text{if} \quad k + \theta \leq j(F_2 - F - 1)\cdot x \leq k + 1,\ k \in \mathbb{Z},\end{cases}\end{aligned}$$

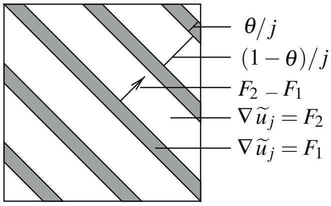

θ/j
$(1-\theta)/j$
$F_2 - F_1$
$\nabla \widetilde{u}_j = F_2$
$\nabla \widetilde{u}_j = F_1$

Fig. 9.2 Function $\widetilde{u}_j$ whose gradient oscillates between F_1 and F_2 on a length scale $1/j$ with volume fractions θ and $1-\theta$

i.e., $\nabla\widetilde{u}_j$ oscillates between the values F_1 and F_2 with frequency j and volume fractions θ and $(1-\theta)$, respectively. If $x\cdot(F_2-F_1)=k/j$ for an integer k, then a change of variables shows that

$$\widetilde{u}_j(x) = F_1\cdot x + \frac{1}{j}\int_0^{j(F_2-F_1)\cdot x} \widetilde{\chi}_\theta(t)\,\mathrm{d}t = F_1\cdot x + \frac{k}{j}(1-\theta) = \overline{F}\cdot x.$$

Hence, the function $\widetilde{u}_j - \widetilde{u}_\mathrm{D}$ vanishes on the lines $L_k = \{x\in\Omega : (F_2-F_1)\cdot x = k/j\}$ and a Poincaré inequality with $\|\nabla u_j\|_{L^\infty(\Omega)} \le 1$ implies that

$$\|\widetilde{u}_j - \widetilde{u}_\mathrm{D}\|_{L^\infty(\Omega)} \le 1/j.$$

To define functions $(u_j)_{j\in\mathbb{N}}$ that satisfy the boundary condition $u_j(x) = \overline{F}\cdot x$ for $x\in\partial\Omega$, we choose nonnegative cut-off functions $\phi_j \in W_0^{1,\infty}(\Omega)$ with $\|\nabla\phi_j\|_{L^\infty(\Omega)} \le j$, $\|\phi\|_{L^\infty(\Omega)} \le 1$, and $\phi_j(x) = 1$ if $\mathrm{dist}(x,\partial\Omega) > 1/j$. We then set

$$u_j = (1-\phi_j)\widetilde{u}_\mathrm{D} + \phi_j\widetilde{u}_j.$$

We have $\|u_j - \widetilde{u}_\mathrm{D}\|_{L^\infty(\Omega)} \le c/j$. Since $\nabla u_j(x)\in\{F_1,F_2\}$ for $\mathrm{dist}(x,\partial\Omega) > 1/j$ and $|\nabla u_j(x)| \le c$ for $\mathrm{dist}(x,\partial\Omega)\le 1/j$, it follows that $I(u_j)\le c/j$ as required. Moreover, we have that $(u_j)_{j\in\mathbb{N}}$ is bounded in $W^{1,\infty}(\Omega)$ and $u_j\to\widetilde{u}_\mathrm{D}$ in $L^\infty(\Omega)$ as $j\to\infty$. □

For affine boundary conditions with $|\overline{F}|<1$ nonuniqueness and nonexistence of solutions can occur.

Proposition 9.2 (Nonuniqueness and nonexistence) *For* $\overline{F} = 0$ *and* $\alpha\ge 0$ *the functional*

$$\widetilde{I}(u) = I(u) + \frac{\alpha}{2}\|u\|^2$$

has no solution if $\alpha > 0$ *and infinitely many solutions if* $\alpha = 0$.

Proof We have $\widetilde{I}\ge 0$. According to Proposition 9.1 there exists a sequence $(u_j)_{j\in\mathbb{N}}\subset W^{1,4}(\Omega)$ with $I(u_j)\to 0$ and $\|u_j\|_{L^\infty(\Omega)}\to 0$ as $j\to\infty$. If $u\in W^{1,4}(\Omega)$ is a minimizer for $\widetilde{I}$, then we have $I(u)=0$ and $(\alpha/2)\|u\|^2 = 0$. In particular, $I(u)=0$

implies that $|\nabla u| = 1$ almost everywhere. If $\alpha > 0$ this leads to a contradiction. If $\alpha = 0$, we note that there exist infinitely many functions $u \in W^{1,\infty}(\Omega)$ with $|\nabla u| = 1$ and $u|_{\partial\Omega} = 0$, e.g., solutions of the Eikonal equations on subsets of Ω. □

The characterization of the infimal energy for affine boundary conditions leads to the rigorous justification of the convexified problem.

Theorem 9.1 (Scalar relaxation) *The convexified functional*

$$I^{cx}(u) = \int_\Omega W^{**}(\nabla u)\,\mathrm{d}x$$

has a minimizer $u \in \mathscr{A}$. The minimizers are exactly the weak limits of infimizing sequences for I in $\mathscr{A}$.

Proof (*sketchted*) For ease of presentation we assume that $\widetilde{u}_{\mathrm{D}}$ is piecewise affine. The existence of a minimizer $u \in \mathscr{A}$ follows from the direct method in the calculus of variations and it remains to construct an infimizing sequence that approximates u. For this, let $(\mathscr{T}_h)_{h>0}$ be a sequence of triangulations of Ω such that $\widetilde{u}_{\mathrm{D}} \in \mathscr{S}^1(\mathscr{T}_h)$ for all $h > 0$. Due to the density of finite element spaces in $W^{1,4}(\Omega)$, there exists a sequence $(u_h)_{h>0} \subset W^{1,4}_{\mathrm{D}}(\Omega)$ with $u_h \in \mathscr{S}^1(\mathscr{T}_h)$ and $u_h \to u$ in $W^{1,4}(\Omega)$ as $h \to 0$. Since I^{cx} is strongly continuous on $W^{1,4}(\Omega)$, there exists for every $\varepsilon > 0$ a number $h_0 > 0$ such that $I^{cx}(u_h) \le I^{cx}(u) + \varepsilon$ for all $0 < h < h_0$. For every $h > 0$ and $T \in \mathscr{T}_h$, we have that $u_h|_T$ is affine and according to Proposition 9.1 there exists a sequence $(u_{T,j})_{j\in\mathbb{N}} \subset W^{1,\infty}(T)$ such that $u_{T,j}|_{\partial T} = u_h|_{\partial T}$, $\|u_h - u_{T,j}\|_{L^\infty(T)} \to 0$, and

$$|T|W^{**}(\nabla u_h|_T) = \lim_{j\to\infty} \int_T W(\nabla u_{T,j})\,\mathrm{d}x.$$

Given $\varepsilon > 0$ we may thus choose for every $T \in \mathscr{T}_h$ a function $u_T \in W^{1,\infty}(T)$ with $u_T|_{\partial T} = u_h|_{\partial T}$ and

$$|T|W^{**}(\nabla u_h|_T) \le \int_T W(\nabla u_T)\,\mathrm{d}x + \varepsilon/|\Omega|.$$

The function $\widetilde{u}_\varepsilon \in L^\infty(\Omega)$, defined by $\widetilde{u}_\varepsilon|_T = u_T$ for all $T \in \mathscr{T}_h$, satisfies $\widetilde{u}_\varepsilon \in W^{1,\infty}(\Omega)$ and

$$I^{cx}(u) \le I^{cx}(u_h) \le I(\widetilde{u}_\varepsilon) \le I^{cx}(u_h) + \varepsilon \le I^{cx}(u) + 2\varepsilon$$

provided $h < h_0$. This proves that the minimizer u for I^{cx} is the weak limit of an infimizing sequence for I. Conversely, if $(u_j)_{j\in\mathbb{N}} \subset \mathscr{A}$ is an infimizing sequence for I with weak limit $u \in \mathscr{A}$, then we have

$$I^{cx}(u) \le \liminf_{j\to\infty} I^{cx}(u_j) \le \liminf_{j\to\infty} I(u_j).$$

According to the first implication there exists an infimizing sequence $(\widetilde{u}_j)_{j\in\mathbb{N}}$ for I such that $I(\widetilde{u}_j) \to I^{cx}(u)$ as $j \to \infty$. Hence we have $I^{cx}(u) = \liminf_{j\to\infty} I(u_j)$. □

Remark 9.1 The theorem implies that for the constant sequence $(I_j)_{j\in\mathbb{N}}$ with $I_j = I$ for all $j \in \mathbb{N}$ we have $I_j \to^{\Gamma} I^{cx}$ as $j \to \infty$ with respect to weak convergence in $W^{1,4}(\Omega)$. Weakly continuous low-order terms can be incorporated in the result.

9.1.2 General Relaxation Result

The discussion of the model problem for a scalar function reveals that the nonexistence of minimizers is related to the nonconvexity of the energy density W and the development of oscillations in infimizing sequences. In particular, infimizing sequences do not converge strongly but satisfy for Lipschitz domains $\omega \subset \mathbb{R}^d$

$$\liminf_{j\to\infty} \frac{1}{|\omega|} \int_{\omega} W(\nabla u_j)\, \mathrm{d}x = W^{**}(\overline{F}),$$

provided that $\nabla u_j \rightharpoonup \overline{F}$ in $L^p(\omega; \mathbb{R}^d)$ as $j \to \infty$. While this is always true for scalar problems, when vectorial problems are considered the right-hand side is only a lower bound which may be strict. This motivates defining the *quasiconvex envelope* for $F \in \mathbb{R}^{m\times d}$ by

$$W^{qc}(F) = \inf_{v\in W_0^{1,\infty}(\omega;\mathbb{R}^m)} \frac{1}{|\omega|} \int_{\omega} W(F + \nabla v)\, \mathrm{d}x.$$

The definition implies that for quasiconvex energy densities, affine functions are solutions of the corresponding minimization problem subject to their own boundary conditions and this yields that corresponding energy functionals are weakly lower semicontinuous on $W^{1,p}(\Omega; \mathbb{R}^m)$ provided a p-growth condition is satisfied. Analogous to the scalar case, one can prove the following theorem.

Theorem 9.2 (General relaxation [8]) *Let $W : \mathbb{R}^{m\times d} \to \mathbb{R}$ be continuous with $c_1(|F|^p - 1) \le W(F) \le c_2(|F|^p + 1)$ for $1 < p < \infty$ and all $F \in \mathbb{R}^{m\times d}$. Given $f \in L^{p'}(\Omega; \mathbb{R}^m)$, $g \in L^{p'}(\Gamma_{\mathrm{N}}; \mathbb{R}^m)$, and $\widetilde{u}_{\mathrm{D}} \in W^{1,p}(\Omega; \mathbb{R}^m)$, the functional*

$$I^{qc}(u) = \int_{\Omega} W^{qc}(\nabla u)\, \mathrm{d}x - \int_{\Omega} f \cdot u\, \mathrm{d}x - \int_{\Gamma_{\mathrm{N}}} g \cdot u\, \mathrm{d}s$$

has a minimizer $u \in \mathscr{A} = \{v \in W^{1,p}(\Omega; \mathbb{R}^m) : v|_{\Gamma_{\mathrm{D}}} = \widetilde{u}_{\mathrm{D}}|_{\Gamma_{\mathrm{D}}}\}$. The minimizers are exactly the weak limits of infimizing sequences for the corresponding functional I. If $d = 1$ or $m = 1$, then we have $W^{qc} = W^{**}$.

The reason for the discrepancy $W^{qc} \neq W^{**}$ is that for given matrices $F_1, F_2 \in \mathbb{R}^{m\times d}$, there exists a nonconstant function $v \in W^{1,\infty}(\Omega; \mathbb{R}^m)$ with $\nabla v \in \{F_1, F_2\}$

almost everywhere in Ω if and only if F_1 and F_2 are compatible in the sense that $\operatorname{rank}(F_2 - F_1) = 1$. This is always satisfied if $d = 1$ or $m = 1$. An efficient characterization of quasiconvex envelopes is an important open problem.

9.1.3 Statistical Description of Oscillations

The discussion of the model problem for a scalar function implies that

$$W(\nabla v_j) \rightharpoonup^* W^{**}(\overline{F}) \tag{9.1}$$

in $L^\infty(\Omega)$ for an infimizing sequence $(v_j)_{j\in\mathbb{N}} \subset W^{1,\infty}(\Omega)$ for I subject to affine boundary conditions described by $\overline{F} \in \mathbb{R}^d$. We also saw that $\nabla v_j \not\to \nabla v$ in general. In particular, we have that

$$\phi(\nabla v_j) \rightharpoonup^* \phi(\overline{F})$$

in $L^\infty(\Omega)$ only holds in general if ϕ is continuous and $\nabla v_j \to \nabla v = \overline{F}$ or if ϕ is affine. For appropriate growth conditions, we have

$$W^{**}(\overline{F}) = \sum_{i=1}^{d+1} \theta_i W(F_i) \tag{9.2}$$

with convex coefficients $(\theta_i)_{i=1,\dots,d+1}$ and $(F_i)_{i=1,\dots,d+1} \subset \mathbb{R}^d$. It appears natural to expect a relation between the infimizing sequence in (9.1) and the convex combination that defines $W^{**}(\overline{F})$ in (9.2). We recall that the gradients of the constructed infimizing sequence $(v_j)_{j\in\mathbb{N}}$ oscillate between the values F_1 and F_2 with volume fractions θ and $1-\theta$. Thus, the family $(\theta_i, F_i)_{i=1,\dots,d+1}$ provides a statistical description of the oscillations in an infimizing sequence $(v_j)_{j\in\mathbb{N}}$. Conversely, it is desirable to extract the convex combination from the infimizing sequence. For this, we notice that we may identify the convex combination with the probability measure

$$\mu = \sum_{i=1}^{d+1} \theta_i \delta_{F_i}$$

with the Dirac measures δ_{F_i}, $i = 1, 2, \dots, d+1$, i.e., for every continuous function $\phi \in C(\mathbb{R}^d)$ we have

$$\langle \mu, \phi \rangle = \sum_{i=1}^{d+1} \theta_i \phi(F_i).$$

We may also identify the sequence $(\nabla v_j)_{j\in\mathbb{N}}$ with the sequence of families of measures $\mu^j : x \mapsto \mu^j_x$ defined for $j \in \mathbb{N}$ and $x \in \Omega$ by

$$\mu^j_x = \delta_{\nabla v_j(x)}.$$

It turns out that within an appropriate space we have the convergence $\mu^j \rightharpoonup^* \mu$ as $j \to \infty$. For the special case of affine boundary conditions, the limiting probability measure is spatially constant. In general, if the weak limit of the infimizing sequence is not affine, then the corresponding measure also depends on $x \in \Omega$. In the special situation that $\nabla v_j \to \nabla v$ strongly in $L^\infty(\Omega)$, then the limiting family of probability measures is given by $\mu_x = \delta_{\nabla v(x)}$ for almost every $x \in \Omega$. The precise mathematical framework for these considerations is provided by a fundamental theorem for which the following definition is required.

Definition 9.1 Let $C_0(\mathbb{R}^{m\times d}) = \{\phi \in C(\mathbb{R}^{m\times d}) : \lim_{|F|\to\infty} \phi(F) = 0\}$ be equipped with the norm $\|\phi\|_{L^\infty(\mathbb{R}^{m\times d})} = \sup_{F\in\mathbb{R}^{m\times d}} |\phi(F)|$. A functional $\mu \in C_0(\mathbb{R}^{m\times d})'$ is called *probability measure* if

$$\langle \mu, \phi\rangle = \int_{\mathbb{R}^{m\times d}} \phi(F)\,\mathrm{d}\mu(F) \geq 0$$

for all $\phi \in C_0(\mathbb{R}^{m\times d})$ with $\phi \geq 0$ and $\langle \mu, 1\rangle = 1$.

This framework enables the following compactness result.

Theorem 9.3 (Fundamental theorem on Young measures [1, 16]) *Let $(z_j)_{j\in\mathbb{N}}$ be a bounded sequence in $L^p(\Omega;\mathbb{R}^{m\times d})$. Then there exists a subsequence $(z_{j_k})_{k\in\mathbb{N}}$ and a family of probability measures, $\mu = (\mu_x)_{x\in\Omega}$ on $\mathbb{R}^{m\times d}$ called* generated Young measure *such that $x \mapsto \langle \phi, \mu_x\rangle$ is measurable for every $\phi \in C_0(\mathbb{R}^{m\times d})$, and if $\psi \in C(\mathbb{R}^{m\times d})$ and the sequence $x \mapsto \psi(z_{j_k}(x))$ is weakly convergent in $L^1(\Omega)$, then the weak limit is $x \mapsto \langle \psi, \mu_x\rangle$.*

Proof (*sketched*) The sequence $(z_j)_{j\in\mathbb{N}}$ is identified with the sequence $(\mu^j)_{j\in\mathbb{N}}$ of probability measures $\mu^j = (\delta_{z_j(x)})_{x\in\Omega}$. This defines a bounded sequence in the space $L^\infty_{w*}(\Omega; C_0(\mathbb{R}^{m\times d})')$ of weakly-* measurable, essentially bounded maps $\Omega \to C_0(\mathbb{R}^{m\times d})'$. By establishing the duality relation

$$L^1(\Omega; C_0(\mathbb{R}^{m\times d}))' = L^\infty_{w*}(\Omega; C_0(\mathbb{R}^{m\times d})')$$

the first assertion is a consequence of the Banach–Alaoglu–Bourbaki theorem. The identification of the limit of $\psi(z_{j_k})$ for $\psi \in C(\mathbb{R}^{m\times d})$ is based on technical approximations. □

A Young measure generated by a weakly convergent sequence $(z_j)_{j\in\mathbb{N}}$ allows us to characterize accumulation points of sequences $(\phi \circ z_j)_{j\in\mathbb{N}}$, e.g., for $W(\nabla v_j)$ with $z_j = \nabla v_j$ for all $j \in \mathbb{N}$. An important open problem related to the characterization of quasiconvexity is the identification of Young measures that are generated by sequences of gradients. In particular, we have that for every $F \in \mathbb{R}^{m\times d}$ there exists a homogeneous Young measure generated by gradients such that

$$W^{qc}(F) = \langle W, \mu\rangle, \qquad \int_{\mathbb{R}^{m\times d}} A\,\mathrm{d}\mu = F.$$

This identity provides a way to reconstruct information about microstructure from the relaxed problem defined by the functional I^{qc}.

Example 9.1 For the sequence $(\widetilde{u}_j)_{j\in\mathbb{N}}$ defined for $j \in \mathbb{N}$ and $x \in \Omega$ by $\widetilde{u}_j(x) = F_1 + \int_0^{(F_2-F_1)\cdot x} \widetilde{\chi}_\theta(js)\,\mathrm{d}s$ we have that the gradients $(\nabla\widetilde{u}_j)_{j\in\mathbb{N}}$ generate the homogeneous Young measure $x \mapsto \mu_x = \theta\delta_{F_1} + (1-\theta)\delta_{F_2}$.

9.2 Direct and Relaxed Numerical Minimization

We discuss in this section the numerical treatment of minimization problems that lead to the development of oscillations, i.e., the construction of infimizing sequences by numerical minimization, and the approximation of the weak limits via the numerical solution of the relaxed functional. We discuss results from [5–7, 11].

9.2.1 A Lower Bound

The definition of the quasiconvex envelope of an energy density motivates investigation of the numerical approximation of minimization problems with affine boundary conditions. While this provides conceptually a way to find the quasiconvex envelope, it turns out that the convergence of the energies is slow in general, due to the formation of oscillations. The following result from [7] is generalized here to the case $p \geq 1$.

Theorem 9.4 (Lower bound [7]) *Let* $d = 2$, $1 \leq p < \infty$, *set* $F_1 = [0, 1]^\top$, $F_2 = [0, -1]^\top$, *and for* $F \in \mathbb{R}^2$ *define*

$$W(F) = \min\{|F - F_1|^p, |F - F_2|^p\}.$$

For a positive integer N *and* $h = 1/N$ *let* $\mathscr{T}_h$ *be the triangulation of* $\Omega = (0, 1)^2$ *depicted in Fig. 9.3. If* $u_h \in \mathscr{S}^1_0(\mathscr{T}_h)$ *is such that for* $0 < \alpha \leq p$, *we have*

$$I(u_h) = \int_\Omega W(\nabla u_h)\,\mathrm{d}x \leq c_1 h^\alpha,$$

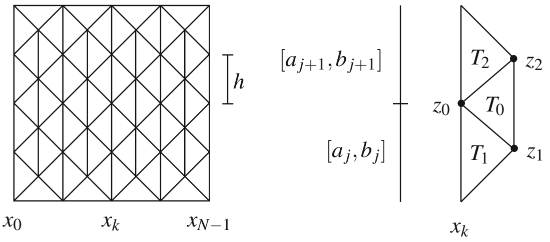


Fig. 9.3 Triangulation for the derivation of the lower bound (*left*) and configuration of *triangles* at the point z_0 considered in the proof of Theorem 9.4 (*right*)

then, provided h is sufficiently small, we have

$$I(u_h) \geq c_1' h^{1-\alpha/p}.$$

In particular, we have $\alpha \leq p/(p+1)$.

Proof (a) Throughout the proof we fix $0 < \delta < 1/5$ independently of h, set $x_k = kh$ for $k = 0, 1, \dots, N-1$, and let K_k be the number of elements $T \in \mathscr{T}_h$ with $T \subset [x_k, x_{k+1}] \times [0, 1]$ and $\min_{\ell=1,2} |\nabla u_h|_T - F_\ell| > \delta$. For every element with this property we have $W(\nabla u_h|_T) \geq \delta^p$ and therefore using $N = h^{-1}$,

$$I(u_h) \geq \sum_{k=0}^{N-1} K_k \delta^p h^2/4 \geq ch \min_{k=0,\dots,N-1} K_k.$$

This implies the asserted bound provided that $K_k \geq ch^{-\alpha/p}$ for $k = 0, 1, \dots, N-1$. To prove this estimate, we fix $0 \leq k \leq N-1$ in the following.
(b) We let $0 \leq J_k \leq N$ and $[a_j, b_j]$, $j = 1, 2, \dots, J_k$, be maximal intervals such that all $T \in \mathscr{T}_h$ with $T \subset [x_k, x_{k+1}] \times [0, 1]$ and an entire edge on $x_k \times [a_j, b_j]$ satisfy either $|\nabla u_h|_T - F_1| \leq \delta$ or $|\nabla u_h|_T - F_2| \leq \delta$. Therefore, we have either $\partial_y u_h \geq 1 - \delta$ on $x_k \times [a_j, b_j]$ or $\partial_y u_h \leq -1 + \delta$ on $x_k \times [a_j, b_j]$ for $j = 1, 2, \dots, J_k$. If $J_k = 0$, then $K_k \geq h^{-1} \geq h^{-\alpha/p}$ so that we may assume $J_k \geq 1$ in the following. If $s \mapsto u_h(x_k, s)$ has a zero $\xi_j \in [a_j, b_j]$ or otherwise with $\xi_j = a_j$, we have

$$|u_h(x_k, s)| \geq \int_{\xi_j}^{s} |\partial_2 u_h(x_k, r)| \,\mathrm{d}r \geq (1 - \delta)|\xi_j - s|.$$

The convexity of $r \mapsto r^{p+1}$ implies that

$$\int_{a_j}^{b_j} |u_h(x_k, s)|^p \,\mathrm{d}s \geq (1 - \delta)^p \int_{a_j}^{b_j} |\xi_j - s|^p \,\mathrm{d}s \geq \frac{(1 - \delta)^p}{p + 1} \frac{1}{2^p} (b - a)^{p+1}.$$

With this we deduce that

$$\int_0^1 |u_h(x_k, s)|^p \,\mathrm{d}s \geq c \sum_{j=1}^{J_k} (b_j - a_j)^{p+1} \geq c \frac{1}{J_k^p} \Big(\sum_{j=1}^{J_k} (b_j - a_j) \Big)^{p+1}.$$

A one-dimensional integration argument, $u_h|_{\partial\Omega} = 0$, and Jensen's inequality prove

$$\int_0^1 |u_h(x_k, s)|^p \,\mathrm{d}s \leq \int_\Omega \big|\partial_1 u_h\big|^p \,\mathrm{d}x.$$

Noting $|\partial_1 u_h(x)|^p \le W(\nabla u_h)$ and $1 - \sum_{j=1}^{J_k}(b_j - a_j) \le K_k h$, the assumption $K_k h \le 1/2$ implies that

$$I(u_h) \ge \int_0^1 |u_h(x_k, s)|^p \,\mathrm{d}s \ge c_2 J_k^{-p}. \tag{9.3}$$

(c) We want to show that $J_k \le K_k + 1$. Since this is true for $J_k = 1$, we assume $J_k \ge 2$ in the following. For two subsequent intervals $[a_j, b_j]$ and $[a_{j+1}, b_{j+1}]$ such that $b_j < a_{j+1}$, there is an element $T_0 \in \mathscr{T}_h$ with $T_0 \subset [x_k, x_{k+1}] \times [0, 1]$ such that an entire edge belongs to $x_k \times [b_j, a_{j+1}]$ and $\left|\nabla u_h|_{T_0} - F_\ell\right| > \delta$, $\ell = 1, 2$. If $b_j = a_{j+1}$, then there exist elements $T_1, T_2 \subset [x_k, x_{k+1}] \times [0, 1]$, such that T_1 has an entire edge on $[a_j, b_j]$ and T_2 has an entire edge on $[a_{j+1}, b_{j+1}]$ and $\left|\nabla u_h|_{T_\ell} - F_\ell\right| \le \delta$ or $\left|\nabla u_h|_{T_\ell} + F_\ell\right| \le \delta$ for $\ell = 1, 2$. For the vertices $z_\ell \in \mathscr{N}_h \cap T_\ell$, $\ell = 1, 2$, not belonging to $x_k \times [0, 1]$ and the vertex $z_0 = (x_k, a_j)$, we have

$$\begin{aligned} u_h(z_1) &= u_h(z_0) + (h/\sqrt{2})\nabla u_h|_{T_1} \cdot [1, 1]^\top, \\ u_h(z_2) &= u_h(z_0) + (h/\sqrt{2})\nabla u_h|_{T_2} \cdot [1, -1]^\top. \end{aligned}$$

On the triangle $T_0 = \mathrm{conv}\{z_0, z_1, z_2\}$ as in the right plot of Fig. 9.3 we thus have

$$\begin{aligned} \partial_y u_h|_{T_0} &= (u_h(z_2) - u_h(z_1))/h \\ &= (\partial_x u_h|_{T_2} - \partial_x u_h|_{T_1})/\sqrt{2} - (\partial_y u_h|_{T_1} + \partial_y u_h|_{T_2})/\sqrt{2} \end{aligned}$$

and $\left|\partial_y u_h|_{T_0}\right| \le 2\sqrt{2}\delta < 4\delta$. This implies that $\left|\nabla u_h|_{T_0} - F_\ell\right| > \delta$ for $\ell = 1, 2$. We have thus shown that between all subsequent intervals $[a_j, b_j]$ and $[a_{j+1}, b_{j+1}]$, $j = 1, 2, \ldots, J_k - 1$, there exists an element T_0 with $\left|\nabla u_h|_{T_0} - F_\ell\right| > \delta$ for $\ell = 1, 2$ and this proves $J_k \le K_k + 1$.

(d) We show that if h is sufficiently small so that $h^{1-\alpha/p} \le (c_1/c_2)^{1/p}/2$, then we may deduce that

$$(c_1/c_2)^{1/p} h^{-\alpha/p} \le K_k + 1.$$

Suppose that the conclusion is wrong. Then

$$(K_k + 1)h < (c_2/c_1)^{1/p} h^{1-\alpha/p} \le 1/2$$

so that (9.3), $J_k \le K_k + 1$, and $c_1 h^\alpha < c_2(K_k + 1)^{-p}$ lead to the contradiction

$$c_1 h^\alpha \ge I(u_h) \ge c_2(K_k + 1)^{-p} > c_1 h^\alpha.$$

This proves the theorem. □

Remark 9.2 On special sequences of triangulations the minimal discrete energy can decay significantly faster with h.

9.2.2 Upper Bounds

To show that the lower bound for the discrete energy minimization is sharp, we consider throughout the energy functional

$$I(u) = \int_\Omega W(\nabla u)\,\mathrm{d}x$$

for $u \in W_0^{1,p}(\Omega)$ and the energy density for $1 \le p < \infty$ and $F \in \mathbb{R}^2$ defined by

$$W(F) = \min\{|F - F_1|^p, |F - F_2|^p\}$$

with $F_1 = [0, 1]^\top$ and $F_2 = [0, -1]^\top$.

Theorem 9.5 (Coarse upper bound) *Given $\Omega \subset \mathbb{R}^2$ and a quasiuniform triangulation $\mathscr{T}_h$ of Ω, we have*

$$\min_{u_h \in \mathscr{S}_0^1(\mathscr{T}_h)} I(u_h) \le c_1 h^{1/2}.$$

Proof We define a function $\widetilde{u} \in W^{1,\infty}(\mathbb{R}^2)$ that is one-periodic in $\widetilde{y}$ by

$$\widetilde{u}(\widetilde{x}, \widetilde{y}) = \begin{cases} \widetilde{y} - k & \text{for} \quad k \in \mathbb{Z} \text{ and } k \le \widetilde{y} \le k + 1/2, \\ 1/2 - (\widetilde{y} - k) & \text{for} \quad k \in \mathbb{Z} \text{ and } k + 1/2 \le \widetilde{y} \le k + 1 \end{cases}$$

and set for $0 < \alpha \le 1$ and $(x, y) \in \Omega$

$$u_\alpha(x, y) = h^\alpha \widetilde{u}(h^{-\alpha}x, h^{-\alpha}y).$$

The function u_α satisfies $\nabla u_\alpha \in \{F_1, F_2\}$ and $\|u_\alpha\|_{L^\infty(\Omega)} \le h^\alpha/2$. For $0 < \beta \le 1$ and $(x, y) \in \Omega$, we define

$$\phi_\beta(x, y) = \min\{h^{-\beta}\,\mathrm{dist}((x, y), \partial\Omega), 1\}$$

which satisfies $\phi_\beta|_{\partial\Omega} = 0$, $\phi_\beta(x, y) = 1$ for $(x, y) \in \Omega$ with $\mathrm{dist}((x, y), \partial\Omega) \ge h^\beta$, and $\|\nabla\psi_\beta\|_{L^\infty(\Omega)} \le ch^{-\beta}$. For the continuous function $u = \psi_\beta u_\alpha$, we have $u|_{\partial\Omega} = 0$, $|\nabla u(x, y)| \le ch^{\alpha-\beta}$ for $(x, y) \in U_{h^\beta}(\partial\Omega) = \{(x, y) \in \Omega : \mathrm{dist}((x, y), \partial\Omega) < h^\beta\}$ and $\nabla u(x, y) \in \{F_1, F_2\}$, otherwise, cf. Fig. 9.4.

The nodal interpolant $u_h = \mathscr{I}_h u$ satisfies $\left|\nabla u_h|_T\right| \le c\|\nabla u\|_{L^\infty(T)}$ for every $T \in \mathscr{T}_h$. We have that $\nabla u_h|_T = F_\ell$ for $\ell \in \{1, 2\}$ and hence $W(\nabla u_h|_T) = 0$ for $T \in \mathscr{T}_h$ if

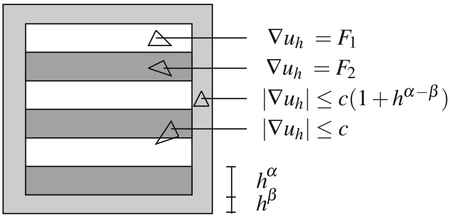


Fig. 9.4 Construction of a finite element function with low energy and typical *triangles*

T does not intersect $U_{h^\beta}(\partial\Omega)$ or a line $L_k = \{(x,y) \in \Omega : y = kh^\alpha\}$ for some $k \in \mathbb{Z}$. The number of such lines in Ω is proportional to $h^{-\alpha}$ and the number of triangles intersecting such a line is bounded by ch^{-1}. Therefore, we deduce, using $|T| \le ch^2$, that

$$I(u_h) \le \int_{U_{h^\beta}} W(\nabla u_h)\,\mathrm{d}x + \sum_{k\in\mathbb{Z}, T\in\mathscr{T}_h : T\cap L_k \ne \emptyset} |T| W(\nabla u_h|_T) \le c(h^\beta + h^\alpha + h^{1-\alpha}).$$

For $\beta = 1$ and $\alpha = 1/2$ we obtain the asserted bound. □

The bound of the previous theorem is sharp for $p = 1$ but can be improved if $p > 1$. To provide a proof for this, we restrict for ease of presentation to the case $\Omega = (0,4) \times (0,1)$ and laminates that are parallel to the x-axis. The main idea is to construct a function that oscillates on a coarse scale in the interior of Ω and on a finer scale in a neighborhood of the boundary, cf. Fig. 9.5. According to Theorem 9.4 the estimate is optimal in the sense that there exists a sequence of triangulations for which it cannot be improved. The following result is based on ideas from [11] and was derived in a more general setting in [5].

Theorem 9.6 (Sharp upper bound) *Let* $p > 1$ *and* $\mathscr{T}_h$ *be a quasiuniform triangulation of* $\Omega = (0,4) \times (0,1)$. *We have*

$$\min_{u_h \in \mathscr{S}^1_0(\mathscr{T}_h)} I(u_h) \le c_1 h^{p/(p+1)}.$$

Proof We define the function $\widetilde{u} \in W^{1,\infty}((0,1)\times\mathbb{R})$ for $(\widetilde{x},\widetilde{y}) \in (0,1)^2$ by

$$\widetilde{u}(\widetilde{x},\widetilde{y}) = \begin{cases} \widetilde{y} & \text{for} \quad 0 \le y \le (1+\widetilde{x})/8, \\ (1+\widetilde{x})/4 - \widetilde{y} & \text{for} \quad (1+\widetilde{x})/8 \le \widetilde{y} \le (3+\widetilde{x})/8, \\ -1/2 + \widetilde{y} & \text{for} \quad (3+\widetilde{x})/8 \le \widetilde{y} \le (5-\widetilde{x})/8, \\ -(1+\widetilde{x})/4 + (1-\widetilde{y}) & \text{for} \quad (5-\widetilde{x})/8 \le \widetilde{y} \le (7-\widetilde{x})/8, \\ \widetilde{y} - 1 & \text{for} \quad (7-\widetilde{x})/8 \le \widetilde{y} \le 1 \end{cases}$$

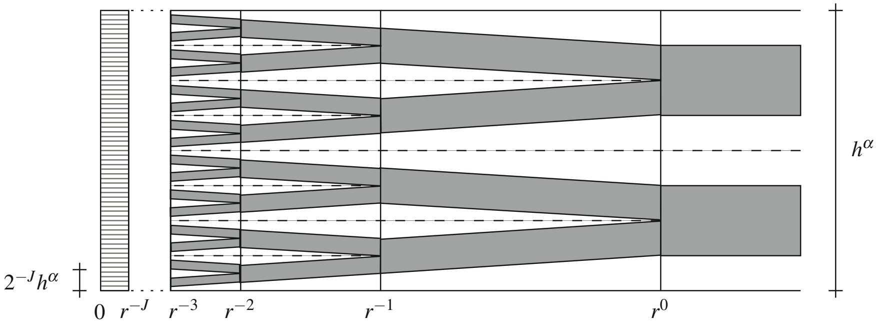


Fig. 9.5 Improved construction of a finite element function with low energy employing multiple scales

and extend $\widetilde{u}$ periodically in $\widetilde{y}$, i.e., $\widetilde{u}(\widetilde{x}, k+\widetilde{y}) = \widetilde{u}(\widetilde{x}, \widetilde{y})$. Notice that $\widetilde{u}(0, y) = \widetilde{u}(1, 2y)$. Given $r > 1$, we let $\psi : [r^{-1}, 1] \to [0, 1]$ be the linear, increasing bijection and set $\widehat{u}(\widehat{x}, \widehat{y}) = \widetilde{u}(\psi(\widehat{x}), \widehat{y})$ for $(\widehat{x}, \widehat{y}) \in [r^{-1}, 1] \times \mathbb{R}$. We then define

$$u(x, y) = \begin{cases} h^\alpha \widehat{u}(1, y/h^\alpha) & \text{for } \ x \in [1, 2], \\ (h^\alpha/2^j)\widehat{u}(r^j x, 2^j y/h^\alpha) & \text{for } \ x \in [r^{-(j+1)}, r^{-j}],\ 0 \le j \le J-1, \\ (h^\alpha/2^J)\widehat{u}(1, 2^J y/h^\alpha) & \text{for } \ x \in [0, r^{-J}]. \end{cases}$$

For $x \in [2, 4]$ we set $u(x, y) = u(4-x, y)$. We assume that $h^{-\alpha} \in \mathbb{Z}$, so that the interval $[0, 1]$ is exactly partitioned into intervals of length h^α and we have $u(x, 0) = u(x, 1) = 0$ for all $x \in [0, 4]$. The construction of the function u is depicted in Fig. 9.5.

With

$$\partial_x u = \pm h^\alpha (r/2)^j \partial_{\widehat{x}} \widehat{u}(r^j x, 2^j y/h^\alpha), \quad \partial_y u = \partial_{\widehat{y}} \widehat{u}(r^j x, 2^j y/h^\alpha).$$

It follows that

$$\nabla u = \begin{cases} F_1 & \text{in the white region,} \\ F_2 + \mathscr{O}\big(h^\alpha (r/2)^j\big)[1, 0]^\top & \text{in the gray region.} \end{cases}$$

The specification of α, r, and J below will guarantee that $h^\alpha (r/2)^j \le c$ for $j = 0, 1, \dots, J$. We define $\phi_\beta(x, y) = \min\{h^{-\beta} x, 1\}$ and set

$$u_h = \mathscr{I}_h[\phi_\beta u].$$

For $(x, y) \in \Omega$ with $x \le h^\delta$ we then have $|\nabla u_h| \le c(1 + 2^{-J} h^{\alpha-\beta})$. The energy $I^j(u_h)$ in the region $[r^{-(j+1)} r^{-j}]$ is the sum of the following contributions:

- For $2^j h^{-\alpha}$ interfaces of length r^{-j} separating regions of constant gradients, we get for $h^{-1} r^{-j}$ many triangles of area h^2 on which $|\nabla u_h| \le c$

$$I^j_{\text{interfaces}} \le c 2^j h^{-\alpha} h^{-1} r^{-j} h^2 = c(2/r)^j h^{1-\alpha}.$$

- In a region of area r^{-j} in which $\nabla u_h = F_2 + \mathscr{O}\big((h^\alpha (r/2)^j\big)[1, 0]^\top$, we obtain

$$I^j_{\text{branching}} \le c r^{-j} \big(h^\alpha (r/2)^j\big)^p = c r^{-j(1-p)} 2^{-jp} = c(r^{p-1}/2^p)^j h^{\alpha p}.$$

We note that in the boundary region $[0, h^\beta] \times [0, 1]$, we have the contribution

$$I_{\text{boundary}} \le c h^\beta (1 + 2^{-J} h^{\alpha-\beta}).$$

Altogether we have

$$I(u_h) \le c h^\beta (1 + 2^{-J} h^{\alpha-\beta}) + c \sum_{j=0}^{J} \big((2/r)^j h^{1-\alpha} + (r^{p-1}/2^p)^{j(1-p)} h^{\alpha p}\big).$$

We choose $2 < r < 2^{p/(p-1)}$, $\beta = p/(p+1)$, $\alpha = 1/(p+1)$, and J as the smallest integer with $J \geq \log_2(h^{\alpha-\beta}) = \log_2 h^{(1-p)/(p+1)})$ so that $2^{-J}h^{\alpha} \leq h^{\beta} = h^{p/(p+1)}$. The choices imply the asserted bound provided $h^{\alpha}(r/2)^J \leq c$. Since $(r/2)^J \leq (2^{1/(p-1)})^J \leq (h^{\alpha-\beta})^{1/(1-p)} = h^{-1/(p+1)} = h^{-\alpha}$ this is guaranteed. □

Remarks 9.3 (i) The result justifies the interpretation of the scale introduced by the discretization as a scale related to a surface energy term.
(ii) The growth parameter $p > 1$ determines the amount of energy stored in interfaces.

9.2.3 Failure of Direct Minimization

The restriction of a coercive and continuous but nonconvex energy functional I to a finite-dimensional subspace leads to the existence of discrete minimizers which define an infimizing sequence as the dimension increases. The problematic analytical properties of the continuous problem are however reflected in the fact that it seems impossible to find global minimizers of the discrete problems due to the occurrence of many local minimizers and large energy barriers between them. The following theorem proves, in a simple model situation, that the number of local minimizers in neighborhoods of decreasing diameter of global minimizers grows exponentially and the separating relative energy barriers increase linearly with the number of degrees of freedom. The statement is a simplified version of a result from [6].

Theorem 9.7 (Local minimizers) *Let $\mathscr{T}_h$ be the uniform partition of $\Omega = (0,1)$ with mesh-size $h = 1/N$ for an even integer N, and for $u_h \in \mathscr{S}^1(\mathscr{T}_h)$ define*

$$I(u_h) = \frac{1}{4}\int_0^1 \left(|u_h'|^2 - 1\right)^2 \mathrm{d}x + \frac{1}{2}\int_0^1 u_h^2 \,\mathrm{d}x.$$

(i) *The minimizers $u_h^{\pm} \in \mathscr{S}^1(\mathscr{T}_h)$ of I satisfy $I(u_h^{\pm}) \leq I_h = h^2/24$ and are given by*

$$u_h^{\pm}|_{T_j}(x) = \pm(-1)^j a(x - x_j)$$

for $x \in T_j = [(j-1)h, jh]$, $x_j = (j-1/2)h$, $j = 1, 2, \ldots, N$, and $a = (1-h^2/12)^{1/2}$, cf. Fig. 9.6.
(ii) *For $u_h \in \{u_h^+, u_h^-\}$ there exist $2^{N/2}$ distinct local minimizers $(u_h^{\ell,*})_{\ell=1,\ldots,2^{N/2}} \subset \mathscr{S}^1(\mathscr{T}_h)$ with $\|u_h^{\ell,*} - u_h\| < 2h$ and $I(u_h^{\ell,*}) \leq 4I_h$, $\ell = 1, 2, \ldots, 2^{N/2}$.*
(iii) *For every continuous path $\varphi : [0,1] \to \mathscr{S}^1(\mathscr{T}_h)$ connecting two of those local minimizers we have $\max_{t\in[0,1]} I(\varphi(t)) \geq 6NI_h$.*

Proof (i) We minimize the energy on every element over affine functions and then assemble the elementwise minimizers to a function in $\mathscr{S}^1(\mathscr{T}_h)$. For the element $T_j = x_j + [-h/2, h/2]$ and $u_h|_{T_j}(x) = a(x - x_j) + b$, we have the energy contribution

$$I_{T_j}(u_h) = \frac{1}{4}\int\limits_{T_j} \big((u_h')^2 - 1\big)^2 \,\mathrm{d}x + \frac{1}{2}\int\limits_{T_j} u_h^2 \,\mathrm{d}x = \frac{h}{4}(a^2-1)^2 + \frac{h^3}{24}a^2 + \frac{h}{2}b^2.$$

A straightforward optimization shows that $a^2 = 1 - h^2/12$ and $b = 0$ and implies that $I_{T_j}(u_h) = (h^3/24)(1 - h^2/12)$. The functions $u_h^\pm \in \mathscr{S}^1(\mathscr{T}_h)$ are obtained by alternating the slopes ±a and these are the only minimizers of I.

(ii) Given $u_h \in \{u_h^\pm\}$ and $1 \le k \le N/2$, we modify u_h on $R_k = T_{2k-1} \cup T_{2k}$ by defining $\widetilde{u}_h \in \mathscr{S}^1(\mathscr{T}_h)$ through $\widetilde{u}_h = u_h$ in $(0,1) \setminus R_k$ and $\widetilde{u}_h'|_{T_{2k-1}} = u_h'|_{T_{2k}}$ and $\widetilde{u}_h'|_{T_{2k}} = u_h'|_{T_{2k-1}}$, cf. Fig. 9.6. To compute $I(\widetilde{u}_h)$, we note that the first part of the energy remains unchanged while the second one is increased. We have

$$\frac{1}{2}\int\limits_{R_k} \widetilde{u}_h^2 - u_h^2 \,\mathrm{d}x = \frac{h^3}{4}a^2, \quad \int\limits_{R_k} (\widetilde{u}_h - u_h)^2 \,\mathrm{d}x = \frac{7h^3}{6}a^2,$$

i.e., $I(\widetilde{u}_h) \le I(u_h) + h^3/4$ and $\|u_h - \widetilde{u}_h\|^2 \le (7/6)h^3$. For every interval R_k, $k = 1, 2, \dots, N/2$, we may either modify u_h as above or leave u_h unchanged which defines $2^{N/2}$ distinct functions $(u_h^\ell)_{\ell=1,\dots,2^{N/2}} \subset \mathscr{S}^1(\mathscr{T}_h)$ with $I(u_h^\ell) \le I_h + (N/2)(h^3/4) = 4I_h$ and $\|u_h - u_h^\ell\|^2 \le (7/12)h^2$. For $\ell = 1, 2, \dots, 2^{N/2}$, we let $u_h^{\ell,*} \in \mathscr{S}^1(\mathscr{T}_h)$ be the minimizer for I within the closure of the set

$$X_h(u_h^\ell) = \{v_h \in \mathscr{S}^1(\mathscr{T}_h) : \operatorname{sign}(v_h') = \operatorname{sign}((u_h^\ell)') \text{ a.e. in } (0,1)\}.$$

We have that $|(u_h^{\ell,*})'| \ge 1/\sqrt{3}$ since otherwise $I(u_h^{\ell,*}) \ge (1/9)h$ which contradicts $I(u_h^{\ell,*}) \le 4I_h$. Since $W''(s) \ge 0$ for $W = (s^2-1)^2/4$ and $s \in \mathbb{R}^2 \setminus B_{1/\sqrt{3}}(0)$, we have that I is strongly convex on the line segment that connects u_h^ℓ and $u_h^{\ell,*}$. Using $DI(u_h^{\ell,*})[v_h] = 0$ for all $v_h \in \mathscr{S}^1(\mathscr{T}_h)$ we verify that

$$\frac{1}{2}\|u_h^{\ell,*} - u_h^\ell\|^2 \le I(u_h^\ell) - I(u_h^{\ell,*}) \le 4I_h - I_h.$$

The triangle inequality yields $\|u_h - u_h^{\ell,*}\| < 2h$.

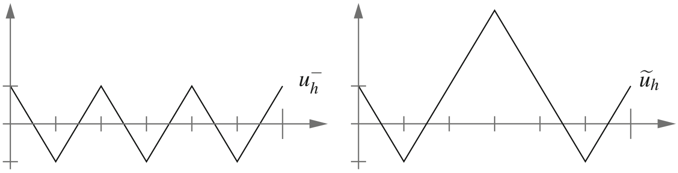


Fig. 9.6 Global discrete minimizer u_h^- (*left*) and modified function $\widetilde{u}_h$ obtained by exchanging the slopes of u_h^- on two adjacent elements (*right*)

(iii) If $\varphi : [0, 1] \to \mathscr{S}^1(\mathscr{T}_h)$ connects two of the above local minimizers, then there exists an element T such that their derivatives have different signs on $T \in \mathscr{T}_h$, e.g., $\varphi(0)'|_T > 0$ and $\varphi(1)'|_T < 0$, and for some $t^* \in [0, 1]$, it follows that $\varphi(t^*)'|_T = 0$ and hence $I(\varphi(t^*)) \geq h/4 = 6h^{-1}I_h$. □

9.2.4 Approximation of the Relaxed Problem

The results on the numerical minimization of the energy functional I with nonconvex energy density show that optimal finite element functions have complicated structures and are, due to the occurrence of local minimizers, difficult to compute. Relaxation theory motivates to discretize the functionals I^{qc} in which W is replaced by its quasiconvex envelope. It can be shown that W^{qc} is continuous and satisfies the same growth conditions as W, so that we assume $W^{qc} \in C(\mathbb{R}^{m \times d})$ with

$$c_1(|F|^p - 1) \leq W^{qc}(F) \leq c_2(|F|^p + 1)$$

for some $1 < p < \infty$, constants $c_1, c_2 > 0$, and all $F \in \mathbb{R}^{m \times d}$. To establish the convergence of finite element minimizers of related energy functionals, it suffices to prove Γ-convergence in the absence of low-order terms and boundary conditions, i.e., to consider

$$I^{qc}(v) = \int_\Omega W^{qc}(\nabla v)\,\mathrm{d}x$$

for $v \in W^{1,p}(\Omega; \mathbb{R}^m)$.

Theorem 9.8 (Convergence) *Let $(\mathscr{T}_h)_{h>0}$ be a sequence of triangulations of Ω and for $h > 0$ and $v \in W^{1,p}(\Omega; \mathbb{R}^m)$, set*

$$I_h^{qc}(v) = \begin{cases} I^{qc}(v) & \text{if } \ v \in \mathscr{S}^1(\mathscr{T}_h)^m, \\ +\infty & \text{if } \ v \in W^{1,p}(\Omega; \mathbb{R}^m) \setminus \mathscr{S}^1(\mathscr{T}_h)^m. \end{cases}$$

We then have $I_h^{qc} \to^\Gamma I^{qc}$ as $h \to 0$ with respect to weak convergence in $W^{1,p}(\Omega; \mathbb{R}^m)$.

Proof We have $I^{qc}(v) \leq I_h^{qc}(v)$ for all $v \in W^{1,p}(\Omega; \mathbb{R}^m)$ and all $h > 0$ and that I^{qc} is weakly lower semi-continuous, so that it suffices to show that for every $v \in W^{1,p}(\Omega; \mathbb{R}^m)$ there exists a sequence $(v_h)_{h>0} \subset W^{1,p}(\Omega; \mathbb{R}^m)$ such that $v_h \in \mathscr{S}^1(\mathscr{T}_h)^m$ for every $h > 0$ and $v_h \rightharpoonup v$ in $W^{1,p}(\Omega; \mathbb{R}^m)$ as $h \to 0$ and

$$I_h^{qc}(v_h) \to I^{qc}(v)$$

as $h \to 0$. Due to the density of the finite element spaces in $W^{1,p}(\Omega; \mathbb{R}^m)$, there exists a sequence of finite element functions $(v_h)_{h>0} \subset W^{1,p}(\Omega; \mathbb{R}^m)$ with $v_h \to v$

in $W^{1,p}(\Omega;\mathbb{R}^m)$ as $h\to 0$. Thus, $\nabla v_h \to \nabla v$ in $L^p(\Omega;\mathbb{R}^{m\times d})$ and for a subsequence we have the pointwise convergence $\nabla v_{h'}(x) \to \nabla v(x)$ for almost every $x\in\Omega$ as $h'\to 0$. Therefore, the continuity of W^{qc} implies $W^{qc}(\nabla v_{h'}) \to W^{qc}(\nabla v)$ almost everywhere in Ω as $h\to 0$. From the growth conditions satisfied by W^{qc}, we deduce that $|W^{qc}(\nabla v_h)| \le c_2(1+|\nabla v_h|^p)$. The generalized dominated convergence theorem thus implies that $I^{qc}(v_{h'}) \to I^{qc}(v)$ as $h'\to 0$. □

Remarks 9.4 (i) Dirichlet boundary conditions are included in the result by considering the restriction of I^{qc} to the space $\widetilde{u}_{\mathrm{D},h} + \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m$ for a sequence of approximations $\widetilde{u}_{\mathrm{D},h} \in \mathscr{S}^1(\mathscr{T}_h)^m$ satisfying $\widetilde{u}_{\mathrm{D},h} \to \widetilde{u}_{\mathrm{D}}$ in $W^{1,p}(\Omega;\mathbb{R}^m)$ as $h\to 0$.
(ii) The same Γ-convergence result can be proved for the functionals I_h that are obtained by restricting the non-quasiconvex functional I to $\mathscr{S}^1(\mathscr{T}_h)^m$, i.e., $I_h \to^{\Gamma} I^{qc}$. The practical minimization of I_h^{qc} is however expected to be significantly simpler than the minimization of I_h, although I_h^{qc} does in general not define a convex minimization problem.

For scalar problems corresponding to $m=1$ or in other special situations, we have $W^{qc} = W^{**}$ with the convex hull W^{**} of W. In this case an error estimate can be proved. We use the fact that for a convex functional $\Phi:\mathbb{R}^{m\times d}\to\mathbb{R}$, we have

$$F\in\partial\Phi^*(S) \iff S\in\partial\Phi(F) \tag{9.4}$$

for all $S,F\in\mathbb{R}^{m\times d}$ which may be interpreted as $[\partial\Phi^*]^{-1} = \partial\Phi$. We note that we also have $\Phi^{***} = \Phi^*$. We say that $\partial\Phi^*$ is *strongly monotone* if

$$c_*|S_1-S_2|^2 \le \langle S_1-S_2, F_1-F_2\rangle$$

for some $c_*>0$ and all $S_1,S_2,F_1,F_2\in\mathbb{R}^{m\times d}$ with $F_\ell\in\partial\Phi^*(S_\ell)$ for $\ell=1,2$, cf. Fig. 9.7 for an illustration.

Theorem 9.9 (Convex relaxation) *Assume that $W^{qc}=W^{**}$, $p=2$, ∂W^* is strongly monotone, and $W^{**}\in C^1(\mathbb{R}^{m\times d})$ with*

$$|DW^{**}(F)| \le c_f'(|F|+1)$$

for all $F\in\mathbb{R}^{m\times d}$. Let $\alpha\ge 0$, $u_0,f\in L^2(\Omega;\mathbb{R}^m)$, and $g\in L^2(\Gamma_{\mathrm{N}};\mathbb{R}^m)$ and suppose that $u\in W^{1,2}_{\mathrm{D}}(\Omega;\mathbb{R}^m)$ and $u_h\in\mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m$ are minimal for

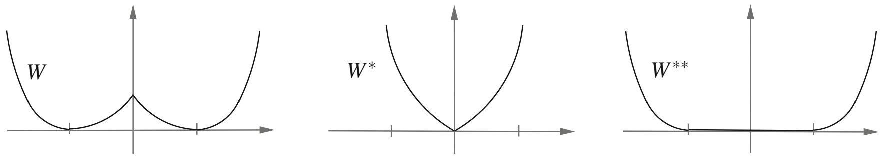


Fig. 9.7 Function $W(F)=\min\{|F-1|^2/2, |F+1|^2/2\}$, its convex hull $W^{**}(F)$ that vanishes in $B_1(0)$ and the conjugate $W^*(S)=W^{***}(S)=|S|^2/2+|S|$ with strongly monotone subdifferential

$$I^{**}(v) = \int_\Omega W^{**}(\nabla v)\,\mathrm{d}x + \frac{\alpha}{2}\|v-u_0\|^2 - \int_\Omega f\cdot v\,\mathrm{d}x - \int_\Omega g\cdot v\,\mathrm{d}s$$

in the sets of all $v \in W^{1,p}_{\mathrm{D}}(\Omega;\mathbb{R}^m)$ and $v \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m$, respectively. With $\sigma = DW^{**}(\nabla u)$ and $\sigma_h = DW^{**}(\nabla u_h)$, we have

$$c_*\|\sigma-\sigma_h\| + \alpha\|u-u_h\| \le \inf_{v_h\in\mathscr{S}^1(\mathscr{T}_h)^m} \big(\|\nabla(u-v_h)\| + \alpha\|u-v_h\|\big).$$

Proof Due to the assumptions on W^{**}, we have that the discrete and continuous minimizers u and u_h satisfy the corresponding Euler–Lagrange equations, i.e.,

$$\int_\Omega DW^{**}(\nabla u)\cdot\nabla v\,\mathrm{d}x + \alpha\int_\Omega (u-u_0)\cdot v\,\mathrm{d}x = \int_\Omega fv\,\mathrm{d}x + \int_{\Gamma_{\mathrm{N}}} gv\,\mathrm{d}s$$

for all $v \in W^{1,p}_{\mathrm{D}}(\Omega;\mathbb{R}^m)$ and

$$\int_\Omega DW^{**}(\nabla u_h)\cdot\nabla v_h\,\mathrm{d}x + \alpha\int_\Omega (u_h-u_0)\cdot v_h\,\mathrm{d}x = \int_\Omega fv_h\,\mathrm{d}x + \int_{\Gamma_{\mathrm{N}}} gv_h\,\mathrm{d}s$$

for all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m$. Since $\mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m \subset W^{1,2}_{\mathrm{D}}(\Omega;\mathbb{R}^m)$, we deduce the Galerkin orthogonality, abbreviating $\sigma = DW^{**}(\nabla u)$ and $\sigma_h = DW^{**}(\nabla u_h)$,

$$\int_\Omega (\sigma-\sigma_h)\cdot\nabla v_h\,\mathrm{d}x + \alpha\int_\Omega (u-u_h)\cdot v_h\,\mathrm{d}x = 0$$

for all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m$. The equivalence (9.4) applied to $\sigma = DW^{**}(\nabla u)$ and $\sigma_h = DW^{**}(\nabla u_h)$ proves $\nabla u \in \partial W^*(\sigma)$ and $\nabla u_h \in \partial W^*(\sigma_h)$, so that the strong monotonicity of ∂W^* yields the estimate

$$c_*|\sigma-\sigma_h|^2 \le (\sigma-\sigma_h)\cdot\nabla(u-u_h)$$

almost everywhere in Ω. This implies that

$$\begin{aligned} c_*\|\sigma-\sigma_h\|^2 + \alpha\|u-u_h\|^2 \le & \int_\Omega (\sigma-\sigma_h)\cdot\nabla(u-u_h)\,\mathrm{d}x \\ & + \alpha\int_\Omega (u-u_h)\cdot(u-u_h)\,\mathrm{d}x. \end{aligned}$$

Galerkin orthogonality allows replacing u_h by an arbitrary function $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^m$ on the right-hand side, and an application of Hölder's inequality leads to the asserted estimate. □

Remarks 9.5 (i) The stresses σ and σ_h are uniquely defined even if u and u_h are non-unique.
(ii) The condition $W^{**} \in C^1(\mathbb{R}^{m\times d})$ and the bound for $|DW^{**}(F)|$ follow from a quadratic growth condition and imply that $W^{***} = W^*$ is strictly convex.
(iii) A corresponding a posteriori error estimate follows analogously.

9.2.5 Iterative Minimization

To find stationary points with low energy for the functional

$$I(u_h) = \int_\Omega W(\nabla u_h)\,\mathrm{d}x$$

with a nonconvex or convex energy density in the set of all $u_h \in \mathscr{S}^1(\mathscr{T}_h)$ with $u_h|_{\Gamma_{\mathrm{D}}} = u_{\mathrm{D}}$, we employ a descent method with line search. We recall that the Armijo–Goldstein criterion guarantees that for $0 < \mu < 1/2$, $u_h \in \mathscr{S}^1(\mathscr{T}_h)$, and $d_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ satisfying

$$(\nabla d_h, \nabla v_h) = -\delta I(u_h)[v_h]$$

for all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$, there exists a number $\alpha' > 0$ such that with $\|\nabla d_h\|^2 = -\delta I(u_h)[d_h]$ we have

$$I(u_h + \alpha d_h) \le I(u_h) - \mu\alpha\|\nabla d_h\|^2.$$

for all $\alpha \in (0, \alpha')$.

Algorithm 9.1 (*Descent method*) Choose $0 < \mu < 1/2$ and $\varepsilon_{\mathrm{stop}} > 0$. Let $u_h^0 \in \mathscr{S}^1(\mathscr{T}_h)$ with $u_h|_{\Gamma_{\mathrm{D}}} = u_{\mathrm{D}}$ and compute a sequence $(u_h^j)_{j=0,1,\ldots} \subset \mathscr{S}^1(\mathscr{T}_h)$ via $u_h^{j+1} = u_h^j + \alpha_j d_h^j$ with $d_h^j \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$, such that

$$(\nabla d_h^j, \nabla v_h) = -\delta I(u_h^j)[v_h] = -\int_\Omega DW(\nabla u_h^j)\cdot\nabla v_h\,\mathrm{d}x$$

for all $v_h \in \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)$ and the maximal $\alpha_j \in \{2^{-\ell} : \ell \in \mathbb{N}_0\}$ satisfying

$$I(u_h^j + \alpha_j d_h^j) \le I(u_h^j) - \mu\alpha_j\delta I(u_h^j)[d_h^j].$$

Stop the iteration if $\|\nabla d_h^j\| \le \varepsilon_{\mathrm{stop}}$.

A MATLAB realization of Algorithm 9.1 is displayed in Fig. 9.8.

```
function energy_minimization(d,red)
[c4n,n4e,Db,Nb] = triang_cube(d); Db = [Db;Nb]; Nb = [];
for j = 1:red
    [c4n,n4e,Db,Nb] = red_refine(c4n,n4e,Db,Nb);
end
[nC,d] = size(c4n); h = 2^(-red);
dNodes = unique(Db); fNodes = setdiff(1:nC,dNodes);
[s,¬,¬,¬] = fe_matrices(c4n,n4e);
F = zeros(d,1); u = c4n*F+h*(rand(nC,1)-.5);
u(dNodes) = c4n(dNodes,:)*F;
sd = zeros(nC,1);
mu = 1/4; norm_corr = 1; eps_stop = 5e-4;
while norm_corr > eps_stop
    alpha = 1;
    [I_0,dI] = energy(c4n,n4e,u);
    sd(fNodes) = -s(fNodes,fNodes)\dI(fNodes);
    [I_alpha,¬] = energy(c4n,n4e,u+alpha*sd);
    armijo = I_alpha-I_0-mu*alpha*dI'*sd;
    while armijo > 0
        alpha = alpha/2;
        [I_alpha,¬] = energy(c4n,n4e,u+alpha*sd);
        armijo = I_alpha-I_0-mu*alpha*dI'*sd;
    end
    u = u+alpha*sd;
    norm_corr = sqrt((alpha*sd)'*s*(alpha*sd))
    show_p1(c4n,n4e,Db,Nb,u); drawnow;
end

function [I,dI] = energy(c4n,n4e,u)
[nC,d] = size(c4n); nE = size(n4e,1);
du = comp_gradient(c4n,n4e,u);
I = 0; dI = zeros(nC,1);
for j = 1:nE
    X_T = [ones(1,d+1);c4n(n4e(j,:),:)'];
    grads_T = X_T\[zeros(1,d);eye(d)];
    vol_T = det(X_T)/factorial(d);
    [W_T,DW_T] = W(du(j,:));
    I = I+vol_T*W_T;
    for k = 1:d+1
        dI(n4e(j,k)) = dI(n4e(j,k))+vol_T*DW_T*grads_T(k,:)';
    end
end

function [val,vec] = W(F)
p = 32; phi = pi/4; F_ref = [cos(phi),sin(phi),0]; d = size(F,2);
F_1 = F_ref(1:d); F_2 = -F_1;
val = (1/p)*norm(F-F_1,2)^(p/2)*norm(F-F_2,2)^(p/2);
vec = (1/2)*norm(F-F_1,2)^(p/2-2)*norm(F-F_2,2)^(p/2)*(F-F_1)+...
  (1/2)*norm(F-F_1,2)^(p/2)*norm(F-F_2,2)^(p/2-2)*(F-F_2);
```

Fig. 9.8 MATLAB routine that realizes a descent method for the minimization of the energy functional $I(u) = \int_\Omega W(\nabla u)\,\mathrm{d}x$ with $W(F) = (1/p)|F-F_1|^{p/2}|F-F_2|^{p/2}$ subject to the affine boundary condition $u(x) = Fx$ for $x \in \Gamma_\mathrm{D}$

Remark 9.6 The algorithm can also be applied to the convexified problem. For that problem it is also desirable to use a Newton iteration. This is however problematic due to the fact that the convex envelope W^{**} is typically only degenerately convex in the sense that $D^2 W^{**}$ may vanish. To avoid related problems stabilizing quadratic terms are often added in the energy minimization.

9.3 Approximation of Semi-convex Envelopes

The relaxed problem defined by the energy functional I^{qc} provides a well-posed reformulation of the original minimization problem and allows us to reconstruct information about the occurrence of microstructures. The essential ingredient is the knowledge of the quasiconvex envelope W^{qc} of the energy density W which can be computed explicitly only in special situations. Its numerical approximation is difficult since no efficient characterization of quasiconvexity is known and the definition of W^{qc} as a minimization problem causes severe numerical difficulties. Upper and lower bounds for W^{qc} are known and these are accessible numerically. We discuss their approximation and consider throughout the case $m = d$ and a continuous function $W : \mathbb{R}^{d\times d} \to \mathbb{R}$ that satisfies

$$c_1(|F|^p - 1) \le W(F) \le c_2(|F|^p + 1)$$

for all $F \in \mathbb{R}^{d\times d}$ with constants $c_1, c_2 > 0$ and a number $p \ge 1$. We follow ideas from [3, 4, 9].

9.3.1 *Upper and Lower Bounds for* W^{qc}

A lower bound for the quasiconvex envelope is defined by the polyconvex envelope, cf. [8].

Definition 9.2 The *polyconvex envelope* W^{pc} of W is the largest polyconvex function $W^{pc} : \mathbb{R}^{d\times d} \to \mathbb{R}$ with $W^{pc} \le W$.

We recall that a polyconvex function is convex in the minors, i.e., in the determinants of square submatrices. This implies that the polyconvex envelope is the largest function $\widehat{W} \circ T$ below W with a convex function $\widehat{W} : \mathbb{R}^{\tau_d} \to \mathbb{R}$ and the minors vector $T : \mathbb{R}^{d\times d} \to \mathbb{R}^{\tau_d}$ given by

$$T(F) = \begin{cases} (F, \det F) \in \mathbb{R}^5 & \text{if} \quad d = 2, \\ (F, \det \widehat{F}_{11}, \dots, \det \widehat{F}_{33}, \det F) \in \mathbb{R}^{19} & \text{if} \quad d = 3, \end{cases}$$

where the matrices $\widehat{F}_{ij}$ are obtained from F by deleting the ith row and jth column. The function W^{pc} can equivalently be characterized through a constrained nonlinear minimization problem, i.e., for $F \in \mathbb{R}^{d\times d}$, we have

$$W^{pc}(F) = \inf\Big\{ \sum_{\ell=1}^{\tau_d+1} \theta_\ell W(A_\ell) : A_\ell \in \mathbb{R}^{d\times d},\ \theta_\ell \ge 0 \sum_{\ell=1}^{\tau_d+1} \theta_\ell = 1 \sum_{\ell=1}^{\tau_d+1} \theta_\ell T(A_\ell) = T(F)\Big\}.$$

If the condition $\sum_{\ell=1}^{\tau_d+1} \theta_\ell T(A_\ell) = T(F)$ is simplified to $\sum_{\ell=1}^{\tau_d+1} \theta_\ell A_\ell = F$, then we obtain the convex envelope $W^{**}(F)$. In particular, we have $W^{**}(F) \le W^{pc}(F)$. Since polyconvexity implies weak lower semicontinuity of integral functionals, it provides a lower bound for the quasiconvex envelope. An upper bound is defined through the rank-one convex envelope.

Definition 9.3 The *rank-one convex envelope* W^{rc} of W is the largest function $W^{rc} : \mathbb{R}^{d\times d} \to \mathbb{R}$ such that $W^{rc} \le W$ and $t \mapsto W(F + t\,ab^\top)$ is convex for all $F \in \mathbb{R}^{d\times d}$ and $a, b \in \mathbb{R}^d$.

To identify W^{rc} as an upper bound for $W^{qc}(F)$, we recall that this quantity is defined by the minimization problem

$$W^{qc}(F) = \inf_{v \in W_0^{1,\infty}(\omega)} |\omega|^{-1} \int_\omega W(F + \nabla v)\,\mathrm{d}x.$$

By the strong continuity of the integral functional, an infimizing sequence has to develop oscillations to decrease the value below $W(F)$. Functions of the form

$$u_\varepsilon(x) = F_1 x + a \int_0^{b\cdot x} \widetilde{\chi}_{(\theta,1)}(\varepsilon^{-1}s)\,\mathrm{d}s$$

for $\varepsilon > 0$ which are appropriately truncated at the boundary oscillate between gradients F_1 and $F_2 = F_1 + ab^\top$ so that

$$W^{qc}(F) \le \theta W(F_1) + (1-\theta) W(F_2)$$

provided that $\theta F_1 + (1-\theta)F_2 = F$. The matrices F_1 and F_2 differ by a rank-one matrix and the process can be repeated by replacing F_1 and F_2 by convex combinations $F_1 = \theta_1 F_{11} + (1-\theta_1)F_{12}$ and $F_2 = \theta_2 F_{21} + (1-\theta_2)F_{22}$, provided that $F_{12} - F_{11}$ and $F_{22} - F_{21}$ are rank-one matrices. This process is illustrated in Fig. 9.9 and motivates defining the *lamination-convex envelope* $W^{\ell c}$ as the pointwise limit $\lim_{k\to\infty} W^k$ of the recursively constructed functions $(W^k)_{k\in\mathbb{N}}$ with $W^0 = W$ and W^{k+1} for $k \ge 0$ and $F \in \mathbb{R}^{d\times d}$ defined by

$$\begin{aligned} W^{k+1}(F) = \inf\big\{ &\theta W^k(F_1) + (1-\theta) W^k(F_2) : \theta \in [0,1],\ F_1, F_2 \in \mathbb{R}^{d\times d}, \\ &\theta F_1 + (1-\theta) F_2 = F,\ \operatorname{rank}(F_2 - F_1) = 1\big\}. \end{aligned}$$

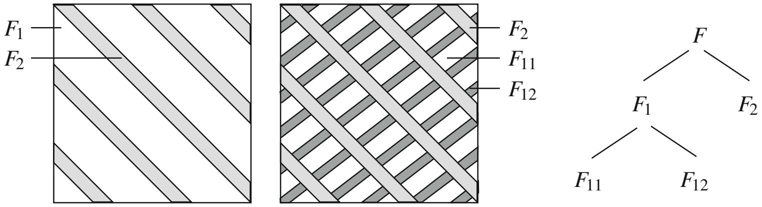


Fig. 9.9 Successive lamination by replacing gradients by convex combinations of rank-one connected matrices

A result in [12] shows that this provides an equivalent characterization of W^{rc}, i.e., $W^{\ell c} = W^{rc}$. In particular, in the iterative process above, the function W is successively convexified along rank-one lines.

The functions W, W^{qc}, W^{pc}, W^{**}, and W^{rc} are related by the inequalities

$$W^{**} \le W^{pc} \le W^{qc} \le W^{rc} \le W.$$

In general all of these inequalities are strict. We remark that only W^{**}, W^{pc}, W^{qc} define weakly lower semicontinuous minimization problems. An important feature of W^{rc} is its physical interpretation that the energy is lowered by iterated laminates.

Example 9.2 ([10]) Given $F_1, F_2 \in \mathbb{R}^{d\times d}$, let

$$W(F) = \min\{W_1(F), W_2(F)\}$$

with $W_j(F) = |F - F_j|^2/2$ for $F \in \mathbb{R}^{d\times d}$ and $j = 1, 2$. Then

$$W^{qc}(F) = \begin{cases} W(F) & \text{for} \quad |W_1(F) - W_2(F)| \ge \lambda/2, \\ W_2(F) - (W_2(F) - W_1(F) + \lambda/2)^2/(2\lambda) & \text{for} \quad |W_1(F) - W_2(F)| \le \lambda/2, \end{cases}$$

where λ is the largest eigenvalue of $(F_2 - F_1)^\top(F_2 - F_1)$. We have $W^{**} \ne W^{pc} = W^{qc} = W^{rc} = W^1$. The identity $W^{**} = W^{rc}$ holds if and only if $\operatorname{rank}(F_2 - F_1) = 1$.

Below, we denote by K_s for $s \ge 0$, the set

$$K_s = \{F \in \mathbb{R}^{d\times d} : |F|_\infty \le s\}$$

with $|F|_\infty = \max_{1\le i,j\le d} |F_{ij}|$. The following assumptions simplify the convergence proofs of the numerical methods for the approximation of W^{pc} and W^{rc}.

Assumption 9.1 (*Convexity and monotonicity*)

(i) There exists a convex function $g : \mathbb{R}^{d\times d} \to \mathbb{R}$ such that $W \ge g$ and $W = g$ in $\mathbb{R}^{d\times d} \setminus K_s$ for some $s > 0$.
(ii) For all $F \in \partial K_s$ and $a, b \in \mathbb{R}^d$, such that $F + t\,ab^\top \notin K_s$ for all $t > 0$, the function $t \mapsto W(F + t\,ab^\top)$ is increasing for $t > 0$.

The first part of the assumption implies that $W^{**} = W^{pc} = W^{qc} = W^{rc} = W$ in $\mathbb{R}^{d\times d} \setminus K_s$. We remark that this is not satisfied in Example 9.2 but the subsequent results can be appropriately modified. The methods described below compute discrete homogeneous Young measures μ_δ^{pc} and μ_δ^{rc}, such that $W^{pc}(F) \approx \langle W, \mu_\delta^{pc}\rangle$ and $W^{rc}(F) \approx \langle W, \mu_\delta^{rc}\rangle$ with a mesh-size $\delta > 0$. The Young measures μ_δ^{rc} will be a gradient Young measures.

9.3.2 Approximation of W^{pc}

To define an approximation of W^{pc} we note that the infimum in the minimization problem that defines $W^{pc}(F)$ remains unchanged if $\tau_d + 1$ is replaced by a number $N \geq \tau_d + 1$, cf. [8]. We then restrict the matrices A_ℓ, $\ell = 1, 2, \ldots, N$, to belong to the grid of nodes $\mathscr{N}_{\delta,r}$ defined for a uniform grid size $\delta > 0$ and a radius $r = N\delta$ for an integer $N > 0$ by

$$\mathscr{N}_{\delta,r} = \delta\mathbb{Z}^{d\times d} \cap K_r,$$

cf. Fig. 9.10. To $\mathscr{N}_{\delta,r}$ we associate the triangulation $\mathscr{T}_{\delta,r}$ of K_r that consists of cubes of edge length δ and with vertices in $\mathscr{N}_{\delta,r}$. The space $\mathscr{S}^1(\mathscr{T}_{\delta,r})$ is the set of all continuous functions on K_r that equal a polynomial of partial degree-one on every cube in $\mathscr{T}_{\delta,r}$. We let $(\varphi_A)_{A\in\mathscr{N}_{\delta,r}}$ be the nodal basis of $\mathscr{S}^1(\mathscr{T}_{\delta,r})$ and $\mathscr{I}_{\delta,r}$ the corresponding nodal interpolation operator.

For $F \in K_r$ and $\mathscr{N} \subset \mathscr{N}_{\delta,r}$ we then consider the optimization problem

$$W_{\mathscr{N}}^{pc}(F) = \inf\Big\{\sum_{A\in\mathscr{N}} \theta_A W(A) : \theta_A \geq 0, \sum_{A\in\mathscr{N}} \theta_A = 1 \sum_{A\in\mathscr{N}} \theta_A T(A) = T(F)\Big\}.$$

In this problem the only degrees of freedom are the convex coefficients $(\theta_A)_{A\in\mathscr{N}}$ that are associated to the fixed nodes $A \in \mathscr{N} \subset \mathbb{R}^{d\times d}$. The problem is thus a linear optimization problem with inequality and equality constraints. To show that the optimization problem $W_{\mathscr{N}}^{pc}(F)$ has a solution, it suffices to construct a feasible

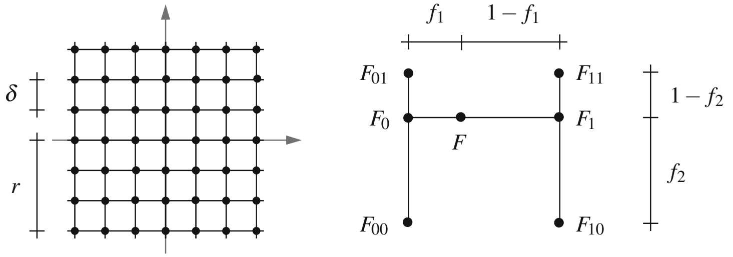


Fig. 9.10 Grid points $\mathscr{N}_{\delta,r} = \delta\mathbb{Z}^{d\times d} \cap K_r$ (*left*); successive decomposition of a matrix $F \in \mathbb{R}^{d\times d} \cap [0,1]^{d\times d}$ along coordinate axes into a convex combination of the vertices of $[0,1]^{d\times d}$ (*right*)

vector $(\theta_A)_{A\in\mathscr{N}}$. This is based on the fact that minors are affine along rank-one lines, e.g., along the coordinate axes.

Lemma 9.1 (Rank-one affinity) *Let $A \in \mathbb{R}^{n\times n}$, $n = 1, 2, \dots, d$, and $1 \le i, j \le n$. For $E_{ij} = e_i e_j^\top \in \mathbb{R}^{n\times n}$ with the canonical basis vectors $e_i, e_j \in \mathbb{R}^n$ and $t \in \mathbb{R}$, we have*

$$\det(A + tE_{ij}) = \det A + (-1)^{i+j}\, t \det \widehat{A}_{ij}.$$

Proof An expansion of the determinant of $A + tE_{ij}$ according to Laplace's rule with respect to the ith row proves the assertion. □

Lemma 9.2 (Local rank-one decomposition) *Let $F \in [0, 1]^{d\times d}$ and $(E_k)_{k=1,\dots,2^{d^2}}$ be the matrices in $\{0, 1\}^{d\times d}$. There exist convex coefficients $(\rho_k)_{k=1,\dots,2^{d^2}}$ with*

$$T(F) = \sum_{i=1}^{2^{d^2}} \rho_k T(E_k).$$

For $1 \le k \le 2^{d^2}$ we have $\rho_k = \varphi_k(F)$.

Proof We identify the matrix F with the vector $(f_k)_{k=1,\dots,d^2}$ via $f_{(i-1)d+j} = (F)_{ij}$, $1 \le i, j \le d$. The decomposition of F is constructed in a hierarchical way in d^2 steps. In the first step we write

$$F = (1 - f_1)F_0 + f_1 F_1,$$

where F_0 and F_1 coincide with F in all components except for the first one in which F_0 vanishes and F_1 has the value 1. Notice that $F_1 - F_0 = e_1 e_1^\top$, so that according to Lemma 9.1 we have

$$T(F) = (1 - f_1)T(F_0) + f_1 T(F_1).$$

In the second step, we write

$$F_0 = (1 - f_2)F_{00} + f_2 F_{01}, \quad F_1 = (1 - f_2)F_{11} + f_2 F_{11},$$

where the matrices F_{00}, F_{01} and F_{10}, F_{11} coincide with F_0 and F_1 except for the second entries, respectively, where F_{00}, F_{10} vanish and F_{01}, F_{11} have the entry 1. The decomposition is sketched in the right plot of Fig. 9.10. Noting that $F_{11} - F_{01} = F_{01} - F_{00} = e_1^\top e_2$ we have

$$T(F) = (1 - f_1)\big[(1 - f_2)T(F_{00}) + f_2 T(F_{01})\big] + f_1\big[(1 - f_2)T(F_{11}) + f_2 T(F_{11})\big].$$

Repeating this procedure we obtain after d^2 steps a decomposition of F with the asserted properties. □

Theorem 9.10 (Approximation) *Let $W \in C^{1,\alpha}(\mathbb{R}^{d\times d})$ and $F \in K_r$ and assume that*

$$W^{pc}(F) = \sum_{\ell=1}^{\tau_d+1} \theta_\ell W(A_\ell)$$

with a feasible family $(\theta_\ell, A_\ell)_{\ell=1}^{\tau_d+1} \subset [0,1] \times K_r$. Then the optimization problem $W^{pc}_{\mathcal{N}_{\delta,r}}(F)$ has a solution, and we have

$$0 \le W^{pc}_{\mathcal{N}_{\delta,r}}(F) - W^{pc}(F) \le c\delta^{1+\alpha}|DW|_{C^{0,\alpha}(K_r)},$$

where $|DW|_{C^{0,\alpha}(K_r)} = \sup_{F_1,F_2\in K_r} |DW(F_1) - DW(F_2)|/|F_1 - F_2|^\alpha$.

Proof Lemma 9.2 implies for $\ell = 1, 2, \ldots, \tau_d + 1$ that

$$T(A_\ell) = \sum_{A\in\mathcal{N}_{\delta,r}} \varphi_A(A_\ell)T(A).$$

Therefore, setting $\theta_A = \sum_{\ell=1}^{\tau_d+1} \theta_\ell\varphi_A(A_\ell)$ we have

$$T(F) = \sum_{A\in\mathcal{N}_{\delta,r}} \theta_A T(A)$$

and $\sum_{A\in\mathcal{N}_{\delta,r}} \theta_A = 1$, cf. Fig. 9.11. Hence, $(\theta_A)_{A\in\mathcal{N}_{\delta,r}}$ defines a feasible vector for the optimization problem, and we have

$$W^{pc}_{\mathcal{N}_{\delta,r}}(F) \le \sum_{A\in\mathcal{N}_{\delta,r}} \theta_A W(A) = \sum_{\ell=1}^{\tau_d+1} \theta_\ell \sum_{A\in\mathcal{N}_{\delta,r}} \varphi_A(A_\ell)W(A) = \sum_{\ell=1}^{\tau_d+1} \theta_\ell \mathcal{I}_\delta W(A_\ell).$$

Since $W^{pc}(F) \le W^{pc}_{\mathcal{N}_{\delta,r}}(F)$, the interpolation estimate

$$\|\mathcal{I}_\delta W - W\|_{L^\infty(K_r(0))} \le c\delta^{1+\alpha}|DW|_{C^{0,\alpha}(K_r(0))}$$

implies the assertion. □

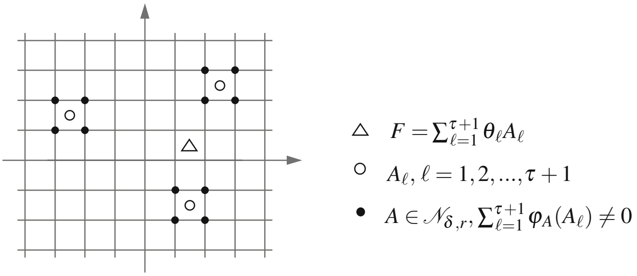

△ $F = \sum_{\ell=1}^{\tau+1} \theta_\ell A_\ell$
○ $A_\ell, \ell = 1,2,\ldots,\tau+1$
● $A \in \mathcal{N}_{\delta,r}, \sum_{\ell=1}^{\tau+1} \varphi_A(A_\ell) \neq 0$

Fig. 9.11 Interpolation of the polyconvex envelope on the grid $\mathcal{N}_{\delta,r}$; only the nodal basis functions corresponding to the *filled dots* contribute to the interpolation

Remarks 9.7 (i) The bound $A_\ell \in K_r$, $\ell = 1, 2, \ldots, \tau_d + 1$, can be guaranteed with growth properties of W if $p > d$. Otherwise it follows from Carathéodory's theorem. (ii) The $C^{1,\alpha}$-regularity of W is only needed in a neighborhood of the region in which $W^{pc} = W$.

9.3.3 Adaptive Computation of $W^{pc}_{\delta,r}$

The discretization of $W^{pc}(F)$ defines an optimization problem with many unknowns and its direct implementation appears difficult and inefficient. In particular, only $\tau_d + 1$ many matrices are needed in the continuous situation and only these matrices have to be approximated locally. The following optimality conditions characterize the relevant nodes $A \in \mathscr{N}$.

Proposition 9.3 (Maximum principle) *The feasible family* $(\theta_A)_{A\in\mathscr{N}} \subset [0, 1]$ *is optimal in* $W^{pc}_{\delta,r}(F)$ *if and only if there exists* $\lambda_{\mathscr{N}} \in \mathbb{R}^{\tau_d}$ *such that*

$$\lambda_{\mathscr{N}} \cdot T(A) - W(A) \le \sum_{A'\in\mathscr{N}} \theta_{A'}\big(\lambda_{\mathscr{N}} \cdot T(A') - W(A')\big) = \lambda_{\mathscr{N}} \cdot T(F) - W^{pc}_{\delta,r}(F)$$

for all $A \in \mathscr{N}$. *If* $\theta_A > 0$ *for* $A \in \mathscr{N}$, *then equality holds.*

Proof The Karush-Kuhn-Tucker optimality conditions state that the feasible vector $(\theta_A)_{A\in\mathscr{N}}$ solves the optimization problem if and only if there exist $\lambda_{\mathscr{N}} \in \mathbb{R}^{\tau_d}$, $\lambda'_{\mathscr{N}} \in \mathbb{R}$, and $(\mu_A)_{A\in\mathscr{N}} \subset \mathbb{R}$ with $\mu_A \le 0$ for all $A \in \mathscr{N}$ such that $\mu_A\theta_A = 0$ for all $A \in \mathscr{N}$ and

$$\sum_{A'\in\mathscr{N}} \rho_{A'} W(A') - \sum_{A'\in\mathscr{N}} \rho_{A'}\lambda_{\mathscr{N}} \cdot T(A') - \lambda'_{\mathscr{N}} \sum_{A'\in\mathscr{N}} \rho_{A'} + \sum_{A'\in\mathscr{N}} \rho_{A'}\mu_{A'} = 0$$

for all $(\rho_A)_{A\in\mathscr{N}} \subset \mathbb{R}$. Given $A \in \mathscr{N}$ set $\rho_A = \theta_A - 1$ and $\rho_{A'} = \theta_{A'}$ for $A' \in \mathscr{N} \setminus \{A\}$. It follows that

$$\lambda_{\mathscr{N}} \cdot T(A) - W(A) \le \sum_{A'\in\mathscr{N}} \theta_{A'}\big(\lambda_{\mathscr{N}} \cdot T(A') - W(A')\big)$$

and the equality of the right-hand side to $\lambda_{\mathscr{N}}(F) - W^{pc}_{\mathscr{N}}(F)$ follows from the definition of the optimization problem. Conversely, assume that $(\theta_A)_{A\in\mathscr{N}}$ is feasible and the condition of the proposition is satisfied. Then, the Karush-Kuhn-Tucker conditions are satisfied with

$$-\mu_A = \sum_{A'\in\mathscr{N}} \theta_{A'}\big(\lambda_{\mathscr{N}} \cdot T(A') - W(A')\big) - \lambda_{\mathscr{N}} \cdot T(A) + W(A) \ge 0,$$

$$\lambda'_{\mathscr{N}} = \Big(\sum_{A' \in \mathscr{N}} W(A') - \sum_{A' \in \mathscr{N}} \lambda_{\mathscr{N}} \cdot T(A') + \sum_{A' \in \mathscr{N}} \mu_{A'} \Big) / \Big(\sum_{A' \in \mathscr{N}} 1 \Big).$$

The complementarity condition $\mu_A \theta_A = 0$ for all $A \in \mathscr{N}$ and the specification of μ_A show that if $\theta_A > 0$, then $\mu_A = 0$ and equality holds. □

Remarks 9.8 (i) Employing the continuous analogue of Proposition 9.3, i.e., that the map $A \mapsto \lambda \cdot T(A) - W^{pc}(A)$ is maximal for $A = F$, i.e., $0 = \lambda \cdot DF(F) - DW^{pc}(F)$, it follows that the quantity $\lambda_{\mathscr{N}} \cdot DT(F)$ approximates $DW^{pc}(F)$.
(ii) If $p > d$ and $\theta_A > 0$, then the estimate $T(A) \cdot \lambda_{\mathscr{N}} \le c_T |\lambda_{\mathscr{N}}| |A|^d$ implies that $c_1 |A|^p - c_1 - c_T |\lambda_{\mathscr{N}}| |A|^d \le |\lambda_{\mathscr{N}}| |T(F)| + c_1$ and shows that $|A| \le r'$ with a number $r' > 0$ that depends on F and $\lambda_{\mathscr{N}}$.

Proposition 9.3 motivates an iterative active set strategy with small subsets of $\mathscr{N}_{\delta,r}$ for the practical solution of $W^{pc}_{\mathscr{N}_{\delta,r}}(F)$ and a local refinement of the grid.

Proposition 9.4 (Active set prediction)
(i) *Let* $\widetilde{\lambda}_{\mathscr{N}} \in \mathbb{R}^{\tau_d}$ *and* $\varepsilon_{\mathrm{AS}} > 0$ *and set*

$$\mathscr{M} = \big\{ A \in \mathscr{N} : \widetilde{\lambda}_{\mathscr{N}} \cdot T(A) - W(A) \ge \max_{A' \in \mathscr{N}} \big(\widetilde{\lambda}_{\mathscr{N}} \cdot T(A') - W(A') \big) - \varepsilon_{\mathrm{AS}} \big\}.$$

If $\sup_{A \in \mathscr{N}_{\delta,r}} |(\lambda_{\mathscr{N}} - \widetilde{\lambda}_{\mathscr{N}}) \cdot T(A)| \le \varepsilon_{\mathrm{AS}}/2$, *then we have* $W^{pc}_{\mathscr{N}}(F) = W^{pc}_{\mathscr{M}}$.
(ii) *Let* $(\theta_A)_{A \in \mathscr{N}_{\delta,r}}$ *be optimal for* $W^{pc}_{\mathscr{N}_{\delta,r}}(F)$ *and define for* $M > 0$

$$\begin{aligned} Z_\delta &= \big\{ A \in \mathscr{N}_{\delta,r} : \lambda_{\mathscr{N}_{\delta,r}} \cdot T(A) - W(A) \le \lambda_{\mathscr{N}_{\delta,r}} \cdot T(F) - W^{pc}_{\delta,r}(F) - \delta M \}, \\ \widehat{Z}_{\delta/2} &= \big\{ A \in \mathscr{N}_{\delta/2,r} : \text{there exists} \, A' \in Z_\delta \text{ with } |A' - A| \le \delta \big\}. \end{aligned}$$

If $(\theta'_A)_{A \in \mathscr{N}_{\delta/2,r}}$ *is optimal for* $W^{pc}_{\mathscr{N}_{\delta/2,r}}(F)$ *with Lagrange-multiplier* $\lambda_{\mathscr{N}_{\delta/2,r}}$, *then we have* $\sum_{A \in \widehat{Z}_{\delta/2}} \theta'_A \le M^{-1} \eta_r$ *for* $\eta_r = |W|_{C^{0,1}(K_r(0))} + |T|_{C^{0,1}(K_r(0))} |\lambda_{\mathscr{N}_{\delta,r}}|$.

Proof (i) If for a solution $(\theta_A)_{A \in \mathscr{N}_{\delta,r}}$ of the optimization problem $W^{pc}_{\mathscr{N}}(F)$ and $A \in \mathscr{N}$, we have $\theta_A > 0$, then we deduce with Proposition 9.3 that

$$\lambda_{\mathscr{N}} \cdot T(A) - W(A) = \lambda_{\mathscr{N}} \cdot T(F) - W^{pc}_{\mathscr{N}_{\delta,r}}(F),$$

and it follows that

$$\begin{aligned} \widetilde{\lambda}_{\mathscr{N}} \cdot T(A) - W(A) &\ge \lambda_{\mathscr{N}} \cdot T(A) - W(A) - \varepsilon_{\mathrm{AS}}/2 \\ &= \lambda_{\mathscr{N}} \cdot T(F) - W^{pc}_{\mathscr{N}_{\delta,r}}(F) - \varepsilon_{\mathrm{AS}}/2 \\ &= \max_{A' \in \mathscr{N}} \lambda_{\mathscr{N}} \cdot T(A') - W(A') - \varepsilon_{\mathrm{AS}}/2 \\ &\ge \max_{A' \in \mathscr{N}} \widetilde{\lambda}_{\mathscr{N}} \cdot T(A') - W(A') - \varepsilon_{\mathrm{AS}}. \end{aligned}$$

This implies that $A \in \mathscr{M}$.

(ii) If $A \in \widehat{Z}_{\delta/2}$ and $A' \in Z_\delta$ are such that $|A - A'| \leq \delta$, then we have

$$W(A) \geq W^{pc}_{\mathcal{N}_{\delta,r}}(F) + \lambda_{\mathcal{N}_{\delta,r}} \cdot \big(T(A) - T(F)\big) + \delta M - \delta \eta_r.$$

For $A \in \mathcal{N}_{\delta/2,r} \setminus \widehat{Z}_{\delta/2}$ the same estimate holds without δM. If $(\theta_A)_{A \in \mathcal{N}_{\delta/2,r}}$ is optimal in the definition of $W^{pc}_{\delta/2,r}(F)$, then

$$W^{pc}_{\delta/2,r}(F) = \sum_{A \in \mathcal{N}_{\delta/2,r}} \theta_A W(A) \geq W^{pc}_{\delta,r}(F) - \delta \eta_r + \sum_{A \in Z'} \theta'_A M \delta$$

and since $W^{pc}_{\delta,r}(F) \geq W^{pc}_{\delta/2,r}(F)$ this proves the second assertion. □

The proposition leads to the following algorithm in which the active set is predicted and iteratively enlarged until the maximum principle holds. Once a solution has been found the grid is refined locally.

Algorithm 9.2 (*Iterative approximation of* $W^{pc}(F)$) Let $r > 0$, $M > 0$, and L be a positive integer and set $\delta = r$, $\widetilde{\lambda} = 0$, $\mathcal{N} = \mathcal{N}_{\delta,r}$, and $\varepsilon_{\mathrm{AS}} = 1$. Until $\delta = 2^{-L} r$ repeat the following steps.

(i) Set

$$\mathcal{N}_{\mathrm{active}} = \big\{A \in \mathcal{N} : \widetilde{\lambda} \cdot T(A) - W(A) \geq \max_{A' \in \mathcal{N}} \big(\widetilde{\lambda} \cdot T(A') - W(A')\big) - \varepsilon_{\mathrm{AS}}\big\}$$

and add further nodes to $\mathcal{N}_{\mathrm{active}}$ to ensure feasibility.

(ii) Solve the optimization problem $W^{pc}_{\mathcal{N}_{\mathrm{active}}}$ and check the maximum principle on $\mathcal{N}$. If this is not satisfied, then increase $\varepsilon_{\mathrm{AS}}$ and continue with (i).

(iii) Refine the grid locally around those nodes $A \in \mathcal{N}$ for which $\lambda_{\mathcal{N}} \cdot T(A) - W(A) > \lambda_{\mathcal{N}} \cdot T(F) - W^{pc}_{\mathcal{N}}(F) - \delta M$ to obtain a new set $\mathcal{N} \subset \mathcal{N}_{\delta/2,r}$, set $\delta = \delta/2$ and $\varepsilon_{\mathrm{AS}} = \delta$, and continue with (i).

Figure 9.12 displays an implementation in MATLAB of the algorithm. The optimality conditions are checked up to a tolerance δ^2. The time-consuming generation of the grid in $\mathbb{R}^{d \times d}$ and its local refinement and coarsening are realized in the C-routines `grid_gen.c` and `loc_grid_ref.c`.

9.3.4 Approximation of W^{rc}

To approximate the upper bound W^{rc} for W^{qc}, we choose as in the approximation of the polyconvex envelope the grid $\mathcal{N}_{\delta,r} = \delta \mathbb{Z}^{d \times d} \cap K_r$ for parameters $\delta > 0$ and $r = N\delta$ for a positive integer N. We employ a set of discrete rank-one matrices that are identified with pairs of vectors in the set

$$\mathscr{R}^1_{\delta,r} = \{(a_\delta, b_\delta) \in \delta \mathbb{Z}^d \times \delta \mathbb{Z}^d : |a_\delta| \leq 2dr,\ 1 - d\delta \leq |b_\delta| \leq 1 + d\delta\}.$$

With this set the iterative algorithm reads as follows.

```
function polyconvexification(d)
F_ref = [pi/4 0 0 0 0 0 0 0 0]; F = F_ref(1:d^2);
r = 4; L = 4; M = 100;
W = @(F)(sum(F.^2,2)-1).^2;
[W_pc,lambda] = multilevel_poly(W,F,r,L,M)

function [W_pc,lambda] = multilevel_poly(W,F,r,L,M)
d = sqrt(size(F,2)); tau_d = (d-1)*d^2+1;
T_F = [F,minors(F)];
delta = r; atoms = grid_gen(delta,r,d);
W_pc = 0; lambda = zeros(tau_d,1);
eps_as = 1; ell = 1;
while ell ≤ L
    W_A = feval(W,atoms); T_A = [atoms,minors(atoms)];
    mp = 0;
    while ¬mp
        active = active_set(lambda,T_F,eps_as,T_A,W_A,delta,d);
        [lambda,W_pc] = lin_prog(T_F,active,T_A,W_A,tau_d);
        mp = max_princ(lambda,T_F,W_pc,delta,T_A,W_A);
        eps_as = eps_as*2;
    end
    if ell < L
        atoms = refine_coarsen(lambda,T_F,W_pc,T_A,W_A,delta,M,d);
        delta = delta/2; eps_as = delta;
    end
    ell = ell+1;
end

function active = active_set(lambda,T_F,eps_as,T_A,W_A,delta,d)
nA = size(T_A,1);
vec = T_A*lambda-W_A;
active = sparse(size(T_A,1),1);
idx_mp = (vec>max(vec)-eps_as);
idx_feas = (max(abs(T_A(:,1:d^2)...
    -ones(nA,1)*T_F(1:d^2)),[],2)≤delta);
active(max(idx_mp,idx_feas)) = 1;

function [lambda,W_pc] = lin_prog(T_F,active,T_A,W_A,tau_d)
idx = find(active); n_active = nnz(idx);
f = W_A(idx)'; A = [T_A(idx,:),ones(n_active,1)]'; b = [T_F,1]';
[¬,W_pc,¬,¬,Lambda] = linprog(f,[],[],A,b,zeros(n_active,1),[]);
lambda = -Lambda.eqlin(1:tau_d);

function mp = max_princ(lambda,T_F,W_pc,delta,T_A,W_A)
vec_A = T_A*lambda-W_A;
mp = ¬(max(vec_A)>T_F*lambda-W_pc+delta^2);

function atoms = refine_coarsen(lambda,T_F,W_pc,T_A,W_A,delta,M,d)
vec = T_A*lambda-W_A;
idx = (vec>T_F*lambda-W_pc-M*delta);
atoms = loc_grid_ref(delta/2,T_A(idx,1:d^2),d);
```

Fig. 9.12 MATLAB realization of the iterative scheme for the computation of the polyconvex envelope

Algorithm 9.3 (*Rank-one convexification*) Given $0 < \delta \le r$ set $W_\delta^0 = \mathscr{I}_{\delta,r} W$ and define for $k = 0, 1, \dots$ the function $W_\delta^{k+1} \in \mathscr{S}^1(\mathscr{T}_{\delta,r})$ for every $F \in \mathscr{N}_{\delta,r}$ by

$$W_\delta^{k+1}(F) = \inf \Big\{ \sum_{\ell \in \mathbb{Z}} \theta_\ell W_\delta^k(F + \delta \ell a_\delta b_\delta^\top) : (a_\delta, b_\delta) \in \mathscr{R}^1_{\delta,r}, \theta_\ell \ge 0, \sum_{\ell \in \mathbb{Z}} \theta_\ell \begin{bmatrix} 1 \\ \ell \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Big\},$$

where W_δ^k is extended by $+\infty$ outside K_r. Stop if $\|W_\delta^{k+1} - W_\delta^k\|_{L^\infty(K_r)} \le \varepsilon_{\text{stop}}$.

Remarks 9.9 (i) The algorithm realizes a successive discrete convexification along rank-one lines. In particular, the infimum in the definition of $W_\delta^{k+1}(F)$ is attained with two points, i.e., there exist $(a_\delta, \beta_\delta) \in \mathscr{R}^1_{\delta,r}$, $\theta \in [0, 1]$, and $\ell_1, \ell_2 \in \mathbb{Z}$ such that $\theta \ell_1 + (1 - \theta \ell_2) = 0$ and

$$W_\delta^{k+1}(F) = \theta W_\delta^k(F + \delta \ell_1 a_\delta b_\delta^\top) + (1 - \theta) W_\delta^k(F + \delta \ell_2 a_\delta b_\delta^\top).$$

(ii) The one-dimensional convexification in the algorithm can be realized as follows: Let $(f_j)_{j=0,\dots,L}$ be a sequence of function values associated with grid points $x_j = x_0 + hj$. Set $g_0 = f_0$ and $f_1 = g_1$. Then for $j \ge 2, 3, \dots$ set $g_j = f_j$ if

$$\frac{g_j - g_{j-1}}{x_j - x_{j-1}} \ge \frac{g_{j-1} - g_{j-2}}{x_{j-1} - x_{j-2}}.$$

Otherwise determine the smallest integer $k \le j - 2$ with

$$\frac{g_j - g_{j-k}}{x_j - x_{j-k}} < \frac{g_{j-k} - g_{j-k-1}}{x_{j-k} - x_{j-k-1}}$$

and replace g_{j-m}, $m = 1, 2, \dots, k$, by

$$g_{j-m} = g_{j-k} + (x_{j-k+m} - x_{j-k}) \frac{g_j - g_{j-k}}{x_j - x_{j-k}}.$$

Points lying above the convex envelope can be eliminated from the list and this allows for a realization of the strategy with complexity $\mathscr{O}(L)$, cf. Fig. 9.13.

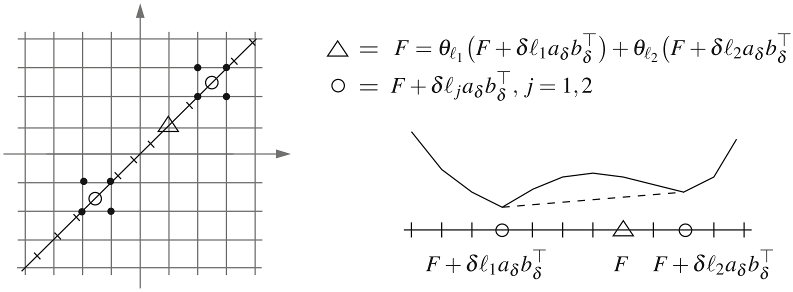

$\triangle = F = \theta_{\ell_1}(F + \delta\ell_1 a_\delta b_\delta^\top) + \theta_{\ell_2}(F + \delta\ell_2 a_\delta b_\delta^\top)$
$\circ = F + \delta\ell_j a_\delta b_\delta^\top, j = 1,2$
$F + \delta\ell_1 a_\delta b_\delta^\top$
F
$F + \delta\ell_2 a_\delta b_\delta^\top$

Fig. 9.13 The iterates in the approximation of the rank-one convex envelope are obtained by discrete convexification along rank-one directions; function values between grid points are obtained by interpolation

Lemma 9.3 (Iterative lamination) *Suppose that Assumption 9.1 holds.*
(i) *For every $F \in K_s$ there exist $F_1, F_2 \in K_s$ and $\theta \in [0,1]$ such that* rank $(F_2 - F_1) = 1$ *and* $W^{k+1}(F) = \theta W^k(F_1) + (1-\theta)W^k(F_2)$.
(ii) *We have* $|W^k|_{C^{0,1}(K_s)} \leq |W|_{C^{0,1}(K_s)}$ *for all* $k \geq 0$.

Proof (i) Due to Assumption 9.1 we have $W = W^k = g$ in $\mathbb{R}^{d\times d} \setminus K_s$ for every $k \geq 0$. Let $\varepsilon > 0$ and $\tilde{\theta} W^k(\widetilde{F}_1) + (1-\tilde{\theta})W^k(\widetilde{F}_2) \leq W^{k+1}(F) + \varepsilon$ for feasible matrices $\widetilde{F}_1, \widetilde{F}_2$. We want to show that we can decrease the value by choosing matrices $F_1, F_2 \in K_s$. With $G = \widetilde{F}_2 - \widetilde{F}_1$, we have $\widetilde{F}_1 = F - (1-\tilde{\theta})G$ and $\widetilde{F}_2 = F + \tilde{\theta}G$. Choosing $\alpha \leq 1 - \tilde{\theta}$ and $\beta \leq \tilde{\theta}$ such that $F_1 = F - \alpha G$ and $F_2 = F + \beta G$ satisfy $F_1, F_2 \in \partial K_s$, we have $W^k(F_\ell) \leq W^k(\widetilde{F}_\ell)$ for $\ell = 1, 2$ and with $\theta = \alpha/(\alpha+\beta)$ that $F = \theta F_1 + (1-\theta)F_2$. Noting that $F_\ell = \mu_\ell \widetilde{F}_1 + (1-\mu_\ell)\widetilde{F}_2$ for $\ell = 1, 2$ and $\mu_\ell \in [0,1]$ and $W^k(F_\ell) \leq \mu_\ell W^k(\widetilde{F}_1) + (1-\mu_\ell)W^k(\widetilde{F}_2)$ shows $\theta W^k(F_1) + (1-\theta)W^k(F_2) \leq \tilde{\theta}W^k(\widetilde{F}_1) + (1-\tilde{\theta})W^k(\widetilde{F}_2)$ which implies the assertion.
(ii) Let $F, G \in K_s$ and assume that $W^{k+1}(F) \leq W^{k+1}(G)$. Let $\theta \in [0,1]$ and $F_1, F_2 \in K_s$ be such that $F = \theta F_1 + (1-\theta)F_2$, $\operatorname{rank}(F_2 - F_1) = 1$, and $W^{k+1}(F) = \theta W^k(F_1) + (1-\theta)W^k(F_2)$. With $G_j = F_j + (G - F)$, $j = 1, 2$, we have that $W^{k+1}(G) \leq \theta W^k(G_1) + (1-\theta)W^k(G_2)$ and therefore

$$\begin{aligned} W^{k+1}(G) - W^{k+1}(F) \leq \theta W^k(G_1) + (1-\theta)W^k(G_2) - \theta W^k(F_1) \\ + (1-\theta)W^k(F_2) \leq |W^k|_{C^{0,1}(K_s)}|F - G|. \end{aligned}$$

If $W^{k+1}(F) \geq W^{k+1}(G)$, then the same estimate follows by interchanging the role of G and F. An inductive argument proves the statement. □

Theorem 9.11 (Approximation of W^{rc}) *Assume that $W \in C^{0,1}(\mathbb{R}^{d\times d})$ such that Assumption 9.1 holds and $W^{rc} = W^K$ for some $K \geq 0$. There exists a constant $c_1 > 0$ such that if δ and r satisfy $s \leq r - c_1\delta$, then we have*

$$\|W_\delta^K - W^{rc}\|_{L^\infty(K_r)} \leq Kc\delta|W|_{C^{0,1}(K_r)}.$$

Proof We show that for all $k \geq 0$, we have

$$\|W_\delta^{k+1} - W^{k+1}\|_{L^\infty(K_r)} \leq c\delta\, |W^k|_{C^{0,1}(K_r)} + \|W_\delta^k - W^k\|_{L^\infty(K_r)}$$

which implies the assertion by incorporating Lemma 9.3 and the interpolation estimate $\|W_\delta^0 - W\|_{L^\infty(K_r))} \leq c\delta|W|_{C^{0,1}(K_r)}$. We consider the case $k = 0$ and abbreviate $W_\delta^0 = W_\delta$ and $W_0 = W$. The general case follows analogously. Let $\widetilde{W}_\delta^1$ be defined for $F \in K_r$ by

$$\widetilde{W}_\delta^1(F) = \inf\Big\{\sum_{\ell\in\mathbb{Z}} \theta_\ell \widetilde{W}(F + \delta\ell a_\delta^\top b_\delta) : (a_\delta, b_\delta) \in \mathscr{R}_{\delta,r}^1, \theta_\ell \geq 0, \sum_{\ell\in\mathbb{Z}} \theta_\ell(1,\ell) = (1,0)\Big\},$$

where $\widetilde{W} = W$ on K_r and $\widetilde{W} = +\infty$ otherwise. Suppose that $\widetilde{W}_\delta^1(F) \leq W_\delta^1(F)$ and let $(a_\delta, b_\delta) \in \mathscr{R}_\delta^1$ and $(\theta_\ell)_{\ell\in\mathbb{Z}}$ be optimal in the definition of $\widetilde{W}_\delta^1(F)$. Then

$$\begin{aligned}|W_\delta^1(F) - \widetilde{W}_\delta^1(F)| &\le \sum_{\ell\in\mathbb{Z}} \theta_\ell \big(W_\delta(F + \delta\ell a_\delta^\top b_\delta) - W(F + \delta\ell a_\delta^\top b_\delta)\big) \\ &\le \|W - W_\delta\|_{L^\infty(K_r)}.\end{aligned}$$

If $\widetilde{W}_\delta^1(F) \ge W_\delta^1(F)$, then the same estimate follows from interchanging the roles of $\widetilde{W}_\delta^1(F)$ and $W_\delta^1(F)$. It remains to bound the difference $|\widetilde{W}_\delta^1(F) - W^1(F)|$. If $F \in K_r \setminus K_s$, we have $W^1(F) = W(F)$ and since $W^1(F) \le \widetilde{W}_\delta^1(F) \le W(F)$ also $\widetilde{W}_\delta^1(F)$. Otherwise, if $F \in K_s$, let $F_1, F_2 \in K_s$, $a, b \in \mathbb{R}^d$, and $\theta \in [0, 1]$, such that $F_2 - F_1 = ab^\top$, $|b| = 1$, $F = \theta F_1 + (1-\theta)F_2$, and

$$W^1(F) = \theta W(F_1) - (1-\theta)W(F_2).$$

We note that $|a| = |F_2 - F_1| \le 2d^{1/2}s$, and for

$$a_\delta = \delta\lfloor a/\delta\rfloor, \quad b_\delta = \delta\lfloor b/\delta\rfloor$$

we deduce that $|a_\delta| \le |a| + d^{1/2}\delta \le 2d^{1/2}r$. and $1 - d^{1/2}\delta \le |b_\delta| \le 1 + d^{1/2}\delta$. We define

$$\ell_1 = -\lfloor(1-\theta)/\delta\rfloor, \quad \ell_2 = \lfloor\theta/\delta\rfloor,$$

and if $\ell_1 = \ell_2 = 0$, then we set $\theta_{\ell_1} = \theta$ and $\theta_{\ell_2} = 1 - \theta$. Otherwise let

$$\theta_{\ell_1} = \frac{\ell_2}{|\ell_1| + \ell_2}, \quad \theta_{\ell_2} = 1 - \theta_{\ell_1}.$$

It follows that $|\theta_{\ell_1} - \theta| \le 2\delta$ if $\delta \le 1/2$ and we have $\theta_{\ell_1}\ell_1 + \theta_{\ell_2}\ell_2 = 0$ and $\theta_{\ell_2} - (1-\theta) = -(\theta_{\ell_1} - \theta)$. We have thus constructed a feasible pair $(a_\delta, b_\delta) \in \mathscr{R}_{\delta,r}^1$ and coefficients $(\theta_\ell)_{\ell\in\mathbb{Z}}$ for the definition of $\widetilde{W}_\delta^1(F)$. Employing $W^1(F) \le \widetilde{W}_\delta^1(F)$ and $F_1 = F - (1-\theta)ab^\top$, $F_2 = F + \theta ab^\top$, a repeated application of the triangle inequality shows that

$$\begin{aligned}|W^1(F) - \widetilde{W}_\delta^1(F)| &\le \theta_{\ell_1} W(F - \delta\ell_1 a_\delta b_\delta^\top) + \theta_{\ell_2} W(F + \delta\ell_2 a_\delta b_\delta^\top) \\ &\quad - \theta W(F_1) - (1-\theta)W(F_2) \\ &= \theta_{\ell_1}\big[W(F - \delta\ell_1 a_\delta b_\delta^\top) - W(F - (1-\theta)ab^\top)\big] \\ &\quad + \theta_{\ell_2}\big[W(F + \delta\ell_2 a_\delta b_\delta^\top) + W(F + \theta ab^\top)\big] \\ &\quad + (\theta_{\ell_1} - \theta)\big[W(F_1) - W(F_2)\big] \\ &\le c\delta|W|_{C^{0,1}(K_r)}.\end{aligned}$$

This implies the assertion. □

Remark 9.10 With Lemma 9.2 one can show that the iterates $(W_\delta^k)_{k=0,1,\dots}$ provide reliable upper bounds for W^{rc}, i.e., $W_\delta^k \ge W^{rc}$ for all $k \ge 0$.

References

1. Ball, J.M.: A version of the fundamental theorem for young measures. In: PDEs and Continuum Models of Phase Transitions (Nice, 1988). Lecture Notes in Physics, vol. 344, pp. 207–215. Springer, Berlin (1989). http://dx.doi.org/10.1007/BFb0024945
2. Ball, J.M., James, R.D.: Fine phase mixtures as minimizers of energy. Arch. Ration. Mech. Anal. **100**(1), 13–52 (1987). http://dx.doi.org/10.1007/BF00281246
3. Bartels, S.: Linear convergence in the approximation of rank-one convex envelopes. M2AN Math. Model. Numer. Anal. **38**(5), 811–820 (2004). http://dx.doi.org/10.1051/m2an:2004040
4. Bartels, S.: Reliable and efficient approximation of polyconvex envelopes. SIAM J. Numer. Anal. **43**(1), 363–385 (2005). http://dx.doi.org/10.1137/S0036142903428840
5. Bartels, S., Prohl, A.: Multiscale resolution in the computation of crystalline microstructure. Numer. Math. **96**(4), 641–660 (2004). http://dx.doi.org/10.1007/s00211-003-0483-8
6. Carstensen, C.: Numerical analysis of microstructure. In: Theory and Numerics of Differential Equations (Durham, 2000), Universitext, pp. 59–126. Springer, Berlin (2001)
7. Chipot, M., Müller, S.: Sharp energy estimates for finite element approximations of nonconvex problems. In: Variations of domain and Free-Boundary Problems in Solid Mechanics (Paris, 1997). Solid Mechanics and Its Applications, vol. 66, pp. 317–325. Kluwer, Dordrecht (1999)
8. Dacorogna, B.: Direct Methods in the Calculus of Variations. Applied Mathematical Sciences, vol. 78, 2nd edn. Springer, New York (2008)
9. Dolzmann, G.K., Walkington, N.J.: Estimates for numerical approximations of rank one convex envelopes. Numer. Math. **85**(4), 647–663 (2000). http://dx.doi.org/10.1007/PL00005395
10. Kohn, R.V.: The relaxation of a double-well energy. Contin. Mech. Thermodyn. **3**(3), 193–236 (1991). http://dx.doi.org/10.1007/BF01135336
11. Kohn, R.V., Müller, S.: Surface energy and microstructure in coherent phase transitions. Commun. Pure Appl. Math. **47**(4), 405–435 (1994). http://dx.doi.org/10.1002/cpa.3160470402
12. Kohn, R.V., Strang, G.: Optimal design and relaxation of variational problems. I–III. Commun. Pure Appl. Math. **39**(3), 113–137, 139–182, 353–377 (1986). http://dx.doi.org/10.1002/cpa.3160390305
13. Luskin, M.: On the computation of crystalline microstructure. In: Acta Numerica, vol. 5, pp. 191–257. Cambridge University Press, Cambridge (1996). http://dx.doi.org/10.1017/S0962492900002658
14. Müller, S.: Variational models for microstructure and phase transitions. In: Calculus of Variations and Geometric Evolution Problems (Cetraro, 1996). Lecture Notes in Mathematics, vol. 1713, pp. 85–210. Springer, Berlin (1999)
15. Pedregal, P.: Variational Methods in Nonlinear Elasticity. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2000)
16. Young, L.C.: Generalized curves and the existence of an attained absolute minimum in the calculus of variations. Comptes Rendus de la Société des Sciences et des Lettres de Varsovie, Classe III **30**, 212–234 (1937)

Chapter 10
Free Discontinuities

10.1 Functions of Bounded Variation

Many important phenomena require the description of physical quantities with discontinuous functions. Although Sobolev functions are not continuous in general, they are too restrictive to admit functions with jumps across lower-dimensional subsets. We introduce in this section the space of functions of bounded variations and discuss its properties. The reader is referred to the textbooks [2, 4, 9] for details.

10.1.1 Derivatives of Discontinuous Functions

Functions in $L^1(\Omega)$ define regular distributions and can be differentiated in the distributional sense, i.e., given $u \in L^1(\Omega)$, its *distributional derivative* is the linear functional $Du : C_c^\infty(\Omega;\mathbb{R}^d) \to \mathbb{R}$ defined by

$$\langle Du, \phi\rangle = -\int_\Omega u \,\operatorname{div}\, \phi \,\mathrm{d}x$$

for every $\phi \in C_c^\infty(\Omega;\mathbb{R}^d)$.

Remark 10.1 For $u \in L^1(\Omega)$ we have $u \in W^{1,1}(\Omega)$ if $Du \in L^1(\Omega;\mathbb{R}^d)$, i.e., if there exists $g \in L^1(\Omega;\mathbb{R}^d)$ such that for all $\phi \in C_c^\infty(\Omega;\mathbb{R}^d)$, we have

$$\langle Du, \phi\rangle = \int_\Omega g \cdot \phi \,\mathrm{d}x.$$

The space $C_0(\Omega;\mathbb{R}^m)$ denotes the completion of the space $C_c^\infty(\Omega;\mathbb{R}^m)$ with respect to the norm $\|v\|_{L^\infty(\Omega)} = \sup_{x\in\Omega} |v(x)|$ for $v \in C_c^\infty(\Omega;\mathbb{R}^m)$, defined through

S. Bartels, *Numerical Methods for Nonlinear Partial Differential Equations*,
Springer Series in Computational Mathematics 47,
DOI 10.1007/978-3-319-13797-1_10

the Euclidean norm on $\mathbb{R}^m$. It is a separable Banach space and its dual is denoted by $\mathscr{M}(\Omega;\mathbb{R}^m)$. The elements in $\mathscr{M}(\Omega;\mathbb{R}^m)$ are through Riesz's representation theorem identified with (*vectorial*) *Radon measures*; and the application of $\mu \in \mathscr{M}(\Omega;\mathbb{R}^m)$ to $v \in C_0(\Omega;\mathbb{R}^m)$ is denoted by

$$\langle \mu, \phi \rangle = \int_\Omega \phi \,\mathrm{d}\mu = \int_\Omega \phi(x)\,\mathrm{d}\mu(x).$$

If $m = 1$, we call μ a scalar Radon measure and write $\mathscr{M}(\Omega)$ for $\mathscr{M}(\Omega;\mathbb{R}^m)$.

Examples 10.1 (i) Every $f \in L^1(\Omega;\mathbb{R}^m)$ defines a Radon measure $\mu_f = f \otimes \mathrm{d}x \in \mathscr{M}(\Omega;\mathbb{R}^m)$ through the Lebesgue integral

$$\langle \mu_f, \phi \rangle = \int_\Omega \phi \cdot f \,\mathrm{d}x$$

for all $\phi \in C_c^\infty(\Omega;\mathbb{R}^m)$. This is a bounded linear functional on $C_0(\Omega;\mathbb{R}^m)$ since

$$\langle \mu_f, \phi \rangle \le \|f\|_{L^1(\Omega)} \|\phi\|_{L^\infty(\Omega)}.$$

(ii) The *Dirac distribution* δ_{x_0} for $x_0 \in \overline{\Omega}$ defines a Radon measure in $\mathscr{M}(\Omega)$ which, for all $\phi \in C_0(\Omega)$ is given by

$$\langle \delta_{x_0}, \phi \rangle = \phi(x_0).$$

(iii) Given a union $C = \cup_{i=1}^{\ell} \Gamma_i$ of Lipschitz continuous curves $\Gamma_i \subset \overline{\Omega}$, $i = 1, 2, \dots, n$, and a function $f \in L^1(C;\mathbb{R}^m)$, we define the Radon measure $\mu_{fC} = f \otimes \mathrm{d}s\lfloor_C$ by setting for $\phi \in C_0(\Omega)$

$$\langle \mu_{fC}, \phi \rangle = \int_C \phi f \,\mathrm{d}s.$$

Definition 10.1 A function $u \in L^1(\Omega)$ is said to be of *bounded variation* if its distributional derivative defines a Radon measure, i.e., if there exists $c \ge 0$ such that

$$\langle Du, \phi \rangle = -\int_\Omega u \;\mathrm{div}\; \phi \,\mathrm{d}x \le c\|\phi\|_{L^\infty(\Omega)}$$

for all $\phi \in C_c^1(\Omega;\mathbb{R}^d)$. The minimal constant $c \ge 0$ with this property is called *total variation* of Du and is given by

$$|Du|(\Omega) = \sup\Big\{ -\int_\Omega u \;\mathrm{div}\; \phi \,\mathrm{d}x : \phi \in C_c^1(\Omega;\mathbb{R}^n),\ \|\phi\|_{L^\infty(\Omega)} \le 1 \Big\}.$$

The space of all such functions is denoted $BV(\Omega)$ and called the *space of functions of bounded variation*. It is equipped with the norm

$$\|u\|_{BV(\Omega)} = \|u\|_{L^1(\Omega)} + |Du|(\Omega)$$

for all $u \in BV(\Omega)$.

We summarize some basic facts about the space $BV(\Omega)$.

Remarks 10.2 (i) The space $BV(\Omega)$ is a nonseparable Banach space.
(ii) We have that $|Du|(\Omega)$ is the operator norm of $Du : C_c^\infty(\Omega; \mathbb{R}^d) \to \mathbb{R}$.
(iii) We have $W^{1,1}(\Omega) \subset BV(\Omega)$ with $\|u\|_{BV(\Omega)} = \|u\|_{W^{1,1}(\Omega)}$ for all $u \in W^{1,1}(\Omega)$.
(vi) We have that $u \in BV(\Omega)$ if and only if there exists $\mu \in \mathscr{M}(\Omega; \mathbb{R}^d)$ such that

$$\int_\Omega u \operatorname{div} \phi \, \mathrm{d}x = -\int_\Omega \phi \, \mathrm{d}\mu$$

for all $\phi \in C_c^1(\Omega; \mathbb{R}^d)$.
(v) If $u \in BV(\Omega)$ and $Du = 0$, then u is constant on every connected component of Ω. Moreover, $u \mapsto |Du|(\Omega)$ is a seminorm on $BV(\Omega)$.
(vi) If $u \in BV(\Omega)$ and $\psi : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous with constant L, then $\psi \circ u \in BV(\Omega)$ with $|D(\psi \circ u)|(\Omega) \le L|Du|(\Omega)$.
(vii) If $\Omega = (a, b) \subset \mathbb{R}^1$ and $u \in BV(\Omega)$, then there exists $\widetilde{u} \in BV(\Omega)$ with $u = \widetilde{u}$ almost everywhere in Ω and

$$|Du|(\Omega) = \sup_{a<x_0<x_1<\cdots<x_n<b} \sum_{j=1}^n |\widetilde{u}(x_j) - \widetilde{u}(x_{j-1})|,$$

where the supremum is over all partitions $a < x_0 < x_1 < \cdots < x_n < b$ with $n \ge 1$.

Typical examples of functions in $BV(\Omega)$ that do not belong to $W^{1,1}(\Omega)$ are functions that are piecewise weakly differentiable and jump across lower-dimensional subsets.

Examples 10.2 (i) For $\Omega = (-1, 1)$ and $u(x) = \operatorname{sign}(x)$, we have

$$\langle Du, \phi\rangle = -\int_{(-1,1)} u\phi' \, \mathrm{d}x = \int_{(0,1)} \phi' \, \mathrm{d}x - \int_{(0,1)} \phi' \, \mathrm{d}x = 2\phi(0)$$

for all $\phi \in C_0^1(\Omega)$, i.e., $Du = 2\delta_0$ and $u \in BV(\Omega)$ with $|Du|(\Omega) = 2$.
(ii) For $\Omega \subset \mathbb{R}^d$ and a Lipschitz domain $E \subset \Omega$, the characteristic function $u = \chi_E$ satisfies

$$\langle D\chi_E, \phi\rangle = -\int_\Omega \chi_E \operatorname{div} \phi \, \mathrm{d}x = -\int_E \operatorname{div} \phi \, \mathrm{d}x = -\int_{\partial E} \phi \cdot n_E \, \mathrm{d}s$$

for all $\phi \in C_0^1(\Omega;\mathbb{R}^d)$ with the outer unit normal n_E on ∂E, i.e., we have $D\chi_E = -n_E \otimes \mathrm{d}s\lfloor_{\partial E}$. This implies that $|D\chi_E|(\Omega) = \mathscr{H}^{d-1}(\partial E)$ is the length or surface area of ∂E for $d = 2$ and $d = 3$, respectively.

Remarks 10.3 (i) If $E \subset \mathbb{R}^d$, then E is said to be of finite perimeter in Ω if $\chi_E \in BV(\Omega)$, and in this case $|D\chi_E|(\Omega)$ is called the *perimeter* of E in Ω. The perimeter generalizes the length or surface area of the boundary of a measurable set $E \cap \Omega$.
(ii) The coarea formula states that the total variation coincides with the integral of the perimeters of the level sets of a function of bounded variation, i.e., we have that

$$|Du|(\Omega) = \int_{-\infty}^{+\infty} |D\chi_{\{u>t\}}|(\Omega)\,\mathrm{d}t.$$

10.1.2 Properties of $BV(\Omega)$

The space $BV(\Omega)$ is an extension of $W^{1,1}(\Omega)$ in the sense that $W^{1,1}(\Omega) \subset BV(\Omega)$ and $\|u\|_{W^{1,1}(\Omega)} = \|u\|_{BV(\Omega)}$ for all $u \in W^{1,1}(\Omega)$. Since the set $C^\infty(\overline{\Omega})$ is dense in $W^{1,1}(\Omega)$, we have that $BV(\Omega)$ is not the closure of $C^\infty(\overline{\Omega})$ with respect to the norm in $BV(\Omega)$. In particular, convergence with respect to the norm in $BV(\Omega)$ or equivalently strong convergence in $BV(\Omega)$ is a notion of convergence that is too restrictive in applications.

Definition 10.2 (i) We say that the sequence $(u_n)_{n\in\mathbb{N}} \subset BV(\Omega)$ *converges intermediately* or *strictly* to $u \in BV(\Omega)$ if $u_n \to u$ in $L^1(\Omega)$ and $|Du_n|(\Omega) \to |Du|(\Omega)$ as $n \to \infty$.
(ii) We say that $(u_n)_{n\in\mathbb{N}} \subset BV(\Omega)$ *converges weakly* to $u \in BV(\Omega)$ if $u_n \to u$ in $L^1(\Omega)$ and $Du_n \rightharpoonup^* Du$ in $\mathscr{M}(\Omega;\mathbb{R}^d)$, i.e., $\langle Du_n, \phi\rangle \to \langle Du, \phi\rangle$ as $n \to \infty$ for every $\phi \in C_0(\Omega;\mathbb{R}^d)$.

Remarks 10.4 (i) The space $BV(\Omega)$ is the dual of a separable Banach space and therefore a natural weak* topology on $BV(\Omega)$ exists. It coincides with the notion of weak convergence introduced in the definition.
(ii) The weak topology in $BV(\Omega)$ in the sense of Banach spaces is difficult to characterize due to the lack of an efficient characterization of $BV(\Omega)'$.
(iii) For $(u_n)_{n\in\mathbb{N}} \subset W^{1,p}(\Omega)$ and $1 < p < \infty$, we have that $u_n \to u$ in $W^{1,p}(\Omega)$ for some $u \in W^{1,p}(\Omega)$ if and only if $u_n \rightharpoonup u$ and $\|u_n\|_{W^{1,p}(\Omega)} \to \|u\|_{W^{1,p}(\Omega)}$ as $n \to \infty$.

Examples 10.3 (i) For $\Omega = (-1, 1)$, let $(u_n)_{n\in\mathbb{N}} \subset BV(\Omega)$ be defined by $u_n(x) = nx$ if $|x| \le 1/n$ and $u_n(x) = \mathrm{sign}(x)$ for $|x| \ge 1/n$, cf. Fig. 10.1. We have that $u_n \to u$ in $L^1(\Omega)$ as $n \to \infty$ for $u(x) = \mathrm{sign}(x)$ for all $x \in \Omega$. Since $|Du_n|(\Omega) = \|\nabla u_n\|_{L^1(\Omega)} = 2$ for all $n \in \mathbb{N}$ and $Du = 2\delta_0$, we have $|Du_n|(\Omega) \to |Du|(\Omega)$; that is, the sequence $(u_n)_{n\in\mathbb{N}}$ converges intermediately to u as $n \to \infty$. Since $(u_n)_{n\in\mathbb{N}} \subset$

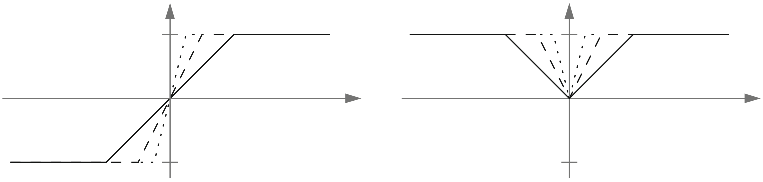


Fig. 10.1 Sequence of functions converging intermediately to $u = \operatorname{sign}$ but not strongly (*left*); sequence of functions converging weakly to $u = 1$ but not intermediately (*right*)

$W^{1,1}(\Omega)$ but $u \notin W^{1,1}(\Omega)$, the sequence does not converge strongly to u.
(ii) For $\Omega = (-1, 1)$ let $(u_n)_{n\in\mathbb{N}}$ be defined by $u_n(x) = n|x|$ if $|x| \le 1/n$ and $u_n(x) = 1$ for $|x| \ge 1/n$, cf. Fig. 10.1. We have that $(u_n)_{n\in\mathbb{N}}$ converges in $L^1(\Omega)$ to the constant function $u = 1$, but $|Du_n|(\Omega) = 2$ and $|Du|(\Omega) = 0$ so that $(u_n)_{n\in\mathbb{N}}$ does not converge intermediately to u. Since $\langle Du_n, \chi_{\{|x|\le 1/m\}}\rangle = 0$ for $m \le n$, it follows that the sequence converges weakly to u.

An important property of the total variation is that it is lower semicontinuous with respect to strong convergence in $L^1(\Omega)$. The following theorem shows that this is equivalent to weak lower semicontinuity in $BV(\Omega)$.

Theorem 10.1 (Weak lower semicontinuity) *If $(u_n)_{n\in\mathbb{N}} \subset BV(\Omega)$ and $u \in L^1(\Omega)$ such that $|Du_n|(\Omega) \le c$ for all $n \in \mathbb{N}$ and $u_n \to u$ in $L^1(\Omega)$, then $u \in BV(\Omega)$ with $|Du|(\Omega) \le \liminf_{n\to\infty} |Du_n|(\Omega)$. Moreover, we have $u_n \rightharpoonup u$ in $BV(\Omega)$ as $n \to \infty$.*

Smooth functions are not dense in $BV(\Omega)$ with respect to strong convergence but with respect to intermediate convergence.

Theorem 10.2 (Approximation by smooth functions) *The spaces $C^\infty(\overline{\Omega})$ and $C^\infty(\Omega) \cap BV(\Omega)$ are dense in $BV(\Omega)$ with respect to intermediate convergence.*

The following compactness result allows us to extract weakly convergent subsequences from bounded sequences in $BV(\Omega)$. This is the crucial difference between the spaces $BV(\Omega)$ and $W^{1,1}(\Omega)$.

Theorem 10.3 (Compactness) *Let $(u_n)_{n\in\mathbb{N}} \subset BV(\Omega)$ be a bounded sequence. Then there exists a subsequence $(u_{n_j})_{j\in\mathbb{N}}$ and $u \in BV(\Omega)$ such that $u_{n_j} \rightharpoonup u$ in $BV(\Omega)$ as $j \to \infty$.*

The most important examples of functions in $BV(\Omega)$ are piecewise regular functions that jump across an interface.

Proposition 10.1 (Piecewise regular functions) *If $\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2$ and Ω_1, Ω_2 are such that $\Omega_1 \cap \Omega_2 = \emptyset$ and $\Sigma = \partial\Omega_1 \cap \partial\Omega_2$ and $u \in L^1(\Omega)$ such that $u|_{\Omega_j} \in W^{1,1}(\Omega_j)$ for $j = 1, 2$, then $u \in BV(\Omega)$ with*

$$Du = \nabla u \otimes \,\mathrm{d}x - [\![un]\!] \otimes \,\mathrm{d}s\lfloor_{\Sigma}$$

with the piecewise defined weak gradient $\nabla u|_{\Omega_j} = \nabla(u|_{\Omega_j})$ *and the jump* $[\![un]\!] = u_{\Omega_1} n_{\Omega_1}\lfloor_{\Sigma} + u_{\Omega_2} n_{\Omega_2}\lfloor_{\Sigma}$ *across* Σ *with the outer unit normals* n_{Ω_j} *to* Ω_j *for* $j = 1, 2$.

Proof For $\phi \in C_c^\infty(\Omega; \mathbb{R}^d)$ a piecewise integration by parts with $\phi|_{\partial\Omega_j \setminus \Sigma} = 0$ for $j = 1, 2$ shows that

$$\begin{aligned}\int_{\Omega} u \,\operatorname{div}\, \phi \,\mathrm{d}x &= \int_{\Omega_1} u \,\operatorname{div}\, \phi \,\mathrm{d}x + \int_{\Omega_2} u \,\operatorname{div},\, \phi \,\mathrm{d}x \\ &= -\int_{\Omega_1} (\nabla u) \cdot \phi \,\mathrm{d}x - \int_{\Omega_2} (\nabla u) \cdot \phi \,\mathrm{d}x + \int_{\partial\Omega_1} u\phi \cdot n_{\Omega_1} \,\mathrm{d}s \\ &\quad + \int_{\partial\Omega_2} u\phi \cdot n_{\Omega_2} \,\mathrm{d}s \\ &= -\int_{\Omega} (\nabla u) \cdot \phi \,\mathrm{d}x + \int_{\Sigma} \phi \cdot [\![un]\!] \,\mathrm{d}s,\end{aligned}$$

which proves the assertion. □

The proposition can be generalized which leads to the following characterization of functions in $BV(\Omega)$.

Theorem 10.4 (Decomposition of Du) *For every* $u \in BV(\Omega)$ *we have*

$$Du = \nabla u \otimes \,\mathrm{d}x - [\![un]\!] \otimes \,\mathrm{d}s|_{S_u} + C_u,$$

where S_u *is a* $(d-1)$*-dimensional jump set,* $\nabla u \in L^1(\Omega)$ *is the weak gradient in the set* $\Omega \setminus S_u$*, and* C_u *either vanishes or is a measure supported on a Cantor set of vanishing* d*-dimensional Hausdorff measure that is zero for sets of finite* $(d-1)$*-dimensional Hausdorff measure. A point* $x \in \Omega$ *belongs to* S_u *if there exists a unit vector* $n \in \mathbb{R}^d$ *and distinct numbers* $a^\pm \in \mathbb{R}$ *such that*

$$\lim_{\varepsilon \to 0} |B_\varepsilon^\pm(x, n) \cap \Omega|^{-1} \int_{B_\varepsilon^\pm(x,n)\cap\Omega} u(y) \,\mathrm{d}y = a^\pm,$$

where $B_\varepsilon^\pm(x, n) = \{y \in B_\varepsilon(x) : \pm(y - x) \cdot n > 0\}$*, cf. Fig.* 10.2.

Some further important properties of $BV(\Omega)$ are listed below.

Remarks 10.5 (i) The embedding $BV(\Omega) \to L^p(\Omega)$ is continuous for $1 \le p \le d/(d-1)$ and compact for $1 \le p < d/(d-1)$.
(ii) We have $\|u\|_{L^p(\Omega)} \le c \,\operatorname{diam}(\Omega)|Du|(\Omega)$ if $u \in BV(\Omega)$ with $\int_\Omega u \,\mathrm{d}x = 0$ and

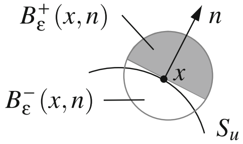


Fig. 10.2 Sets $B_\varepsilon^\pm(x,n)$ for a point $x \in S_u$ where the function u jumps from the value a^- to the value a^+ in the direction of n

$1 \le p \le d/(d-1)$.
(iii) There exists a linear operator $\mathrm{tr} : BV(\Omega) \to L^1(\partial\Omega)$ such that $\mathrm{tr}(u) = u|_{\partial\Omega}$ for all $u \in BV(\Omega) \cap C(\overline{\Omega})$. Moreover, we have the integration by parts formula

$$\int_\Omega \phi Du = -\int_\Omega u \operatorname{div} \phi \,\mathrm{d}x + \int_{\partial\Omega} \mathrm{tr}(u)\phi \cdot n \,\mathrm{d}s$$

for all $u \in BV(\Omega)$ and all $\phi \in C^1(\overline{\Omega}; \mathbb{R}^d)$. The operator tr is continuous with respect to intermediate convergence in $BV(\Omega)$. It is not continuous with respect to weak convergence in $BV(\Omega)$; for example for $(u_n)_{n\in\mathbb{N}} \subset BV(0,1)$ defined through $u_n(x) = nx$ for $x \le 1/n$ and $u(x) = 1$ for $x \ge 1/n$, we have $u_n \rightharpoonup u$ with $u \equiv 1$ but $u_n(0) = 0$ for all $n \in \mathbb{N}$.

10.1.3 A Variational Model Problem on $BV(\Omega)$

To understand the finite element approximation of variational problems involving total variation, we consider, for given $g \in L^2(\Omega)$ and $\alpha > 0$, minimizing the functional

$$I(u) = |Du|(\Omega) + \frac{\alpha}{2}\int_\Omega (u-g)^2 \,\mathrm{d}x$$

as defined for $u \in BV(\Omega) \cap L^2(\Omega)$. By the density of smooth functions we may choose a bounded infimizing sequence $(u_n)_{n\in\mathbb{N}} \subset W^{1,1}(\Omega) \cap L^2(\Omega)$. Due to the lack of reflexivity or more generally an existing separable predual space, we cannot extract a weakly convergent subsequence in $W^{1,1}(\Omega)$. A weak limit of a subsequence exists in the space $BV(\Omega) \cap L^2(\Omega)$.

Theorem 10.5 (Existence) *There exists a minimizer $u \in BV(\Omega) \cap L^2(\Omega)$ for I.*

Proof The functional I is bounded from below and the set of admissible functions is nonempty, and hence there exists a bounded infimizing sequence $(u_n)_{n\in\mathbb{N}} \subset BV(\Omega) \cap L^2(\Omega)$. Theorem 10.3 guarantees the existence of a weakly convergent subsequence $(u_{n_j})_{j\in\mathbb{N}}$ with weak limit $u \in BV(\Omega)$ and Theorem 10.1 implies $I(u) \le \liminf_{j\to\infty} I(u_{n_j})$, i.e., u is a minimizer for I. □

Remark 10.6 The existence of solutions subject to Dirichlet boundary conditions $u|_{\partial\Omega} = u_{\mathrm{D}}$ for $u_{\mathrm{D}} \in L^1(\partial\Omega)$ is difficult to establish due to the lack of weak continuity of the trace operator.

The following stability result implies the uniqueness of minimizers.

Theorem 10.6 (Stability and uniqueness) *For $g_1, g_2 \in L^2(\Omega)$ let the functions $u_1, u_2 \in BV(\Omega) \cap L^2(\Omega)$ be minimizers of I with g replaced by g_1 and g_2, respectively. We then have*

$$\|u_1 - u_2\| \le \|g_1 - g_2\|.$$

In particular, minimizers are uniquely defined.

Proof We define the convex functionals $F : BV(\Omega) \to \mathbb{R}$ and $G_\ell : L^2(\Omega) \to \mathbb{R}$, $\ell = 1, 2$, by

$$F(u) = |Du|(\Omega), \quad G_\ell(u) = (\alpha/2)\|u - g_\ell\|^2$$

and set $I_\ell = F + G_\ell$. We extend F to $L^2(\Omega)$ with the value $+\infty$, and note that G_ℓ is Fréchet differentiable with

$$\delta G_\ell(u)[v] = \alpha(u - g_\ell, v)$$

for all $v \in L^2(\Omega)$. Since F is convex, we have that its subdifferential is monotone, i.e., for $\mu_\ell \in \partial F(u_\ell)$, $\ell = 1, 2$, we have

$$(\mu_2 - \mu_1, u_2 - u_1) \ge 0.$$

Noting that $0 \in \partial I_\ell(u_\ell)$ we deduce that $-\delta G_\ell(u_\ell) \in \partial F(u_\ell)$ for $\ell = 1, 2$, and therefore

$$\big(- \alpha(u_2 - g_2) + \alpha(u_1 - g_1), u_2 - u_1\big) \ge 0.$$

This implies that

$$\|u_2 - u_1\|^2 \le (u_2 - u_1, g_2 - g_1)$$

and an application of Hölder's inequality proves the asserted bound. □

Due to a monotonicity property of the total variation, a maximum principle holds for the minimization problem.

Proposition 10.2 (Maximum principle) *If $g \in L^\infty(\Omega)$, then the minimizer $u \in BV(\Omega) \cap L^2(\Omega)$ for I satisfies $u \in L^\infty(\Omega)$ with $\|u\|_{L^\infty(\Omega)} \le \|g\|_{L^\infty(\Omega)}$.*

Proof Assume that $g(x) \le \overline{g}$ for almost every $x \in \Omega$ and given the minimizer $u \in BV(\Omega) \cap L^2(\Omega)$ for I, define $\widetilde{u}(x) = \min\{u(x), \overline{g}\}$ for $x \in \Omega$. According to Remark 10.2 we have $\widetilde{u} \in BV(\Omega)$ with $|D\widetilde{u}|(\Omega) \le |Du|(\Omega)$. Since also $\|\widetilde{u} - g\| \le \|u - g\|$, we deduce that $I(\widetilde{u}) \le I(u)$. This implies $u = \widetilde{u}$ and $u \le \overline{g}$. The same argument shows that $u \ge \underline{g}$ if $g(x) \ge \underline{g}$ for almost every $x \in \Omega$. Therefore $u \in L^\infty(\Omega)$ with the asserted bound. □

Useful information about the minimization of I is contained in the related dual problem. To identify it, we note that by a completion of $C_c^\infty(\Omega;\mathbb{R}^d)$ with respect to the norm $\|p\|_{H(\text{div};\Omega)} = \|p\| + \|\operatorname{div} p\|$, the total variation $|Du|(\Omega)$ of a function $u \in BV(\Omega) \cap L^2(\Omega)$ can equivalently be characterized as

$$|Du|(\Omega) = \sup\Big\{ -\int_\Omega u \operatorname{div} p \,\mathrm{d}x : p \in H_N(\text{div};\Omega),\ |p| \le 1 \text{ in } \Omega\Big\},$$

where

$$H_N(\text{div};\Omega) = \{p \in L^2(\Omega;\mathbb{R}^d) : \operatorname{div} p \in L^2(\Omega),\ p\cdot n|_{\partial\Omega} = 0\}.$$

For the minimization problem defined through I, we thus have with the indicator functional $I_{K_1(0)}$ of the set

$$K_1(0) = \{p \in L^2(\Omega;\mathbb{R}^d) : |p| \le 1 \text{ almost everywhere in } \Omega\}$$

that

$$\begin{aligned}\inf_{u\in BV\cap L^2} I(u) &= \inf_{u\in BV\cap L^2} |Du|(\Omega) + \frac{\alpha}{2}\|u-g\|^2\\ &= \inf_{u\in BV\cap L^2}\ \sup_{p\in H_N(\text{div})} \Big(-\int_\Omega u \operatorname{div} p\,\mathrm{d}x + \frac{\alpha}{2}\|u-g\|^2 - I_{K_1(0)}(p)\Big).\end{aligned}$$

This defines a saddle point problem with unknowns u and p. The dual problem is obtained by eliminating u. For this we assume that the order of the infimum and supremum can be interchanged, i.e.,

$$\inf_{u\in BV\cap L^2} I(u) = \sup_{p\in H_N(\text{div})}\ \inf_{u\in BV\cap L^2} \Big(-\int_\Omega u \operatorname{div} p\,\mathrm{d}x + \tfrac{\alpha}{2}\|u-g\|^2 - I_{K_1(0)}(p)\Big).$$

A direct calculation shows that the solution u of the inner minimization problem is for $p \in H_N(\text{div};\Omega)$ given by

$$u = g + \alpha^{-1} \operatorname{div} p,$$

and thus

$$\inf_{u\in BV\cap L^2} I(u) = \sup_{p\in H_N(\text{div})} -\frac{1}{2\alpha}\|\operatorname{div} p + \alpha g\|^2 + \frac{1}{2\alpha}\|\alpha g\|^2 - I_{K_1(0)}(p).$$

The maximization problem defined by the right-hand side is the dual problem. The heuristic interchange of the infimum and the supremum can be rigorously justified and leads to the following result.

Proposition 10.3 (Strong duality) *For $p \in H_N(\operatorname{div}; \Omega)$ define*

$$D(p) = -\frac{1}{2\alpha}\| \operatorname{div} p + \alpha g\|^2 + \frac{\alpha}{2}\|g\|^2 - I_{K_1(0)}(p).$$

We have

$$\inf_{u \in BV(\Omega)\cap L^2(\Omega)} I(u) = \sup_{p \in H_N(\operatorname{div};\Omega)} D(p).$$

Moreover, there exists a solution $p \in H_N(\operatorname{div}; \Omega)$ that maximizes the functional D.

Proof The reader is referred to [12] for a proof of the result which is established by showing that I is the Fenchel dual of D in the sense of [11]. □

Remark 10.7 Exchanging the order of the infimum and supremum always leads to the weak duality principle $\inf_u I(u) \geq \sup_p D(p)$.

Proposition 10.4 *The unique solution $u \in BV(\Omega) \cap L^2(\Omega)$ of the minimization problem defined by I and every solution $p \in H_N(\operatorname{div}; \Omega)$ of the maximization problem defined by D correspond to a saddle point for the functional*

$$L(u, p) = -\int_\Omega u \operatorname{div} p \,\mathrm{d}x + \frac{\alpha}{2}\|u - g\|^2 - I_{K_1(0)}(p)$$

and are related by

$$\operatorname{div} p = \alpha(u - g), \quad Du \in \partial I_{K_1(0)}(p),$$

where the inclusion is understood as

$$-\big(u, \operatorname{div}(q - p)\big) \leq 0$$

for all $q \in H_N(\operatorname{div}; \Omega) \cap K_1(0)$.

Proof The proof follows from standard arguments in convex optimization, cf., e.g., [11]. □

Remarks 10.8 (i) The inclusion $Du \in \partial I_{K_1(0)}(p)$ is formally equivalent to $p \in \partial|Du|$. In particular, we have $p = \nabla u/|\nabla u|$ in regions where $\nabla u \neq 0$.
(ii) In the case of Dirichlet boundary conditions on $\partial\Omega$, the space $H_N(\operatorname{div}; \Omega)$ is replaced by $H(\operatorname{div}; \Omega) = \{p \in L^2(\Omega; \mathbb{R}^d) : \operatorname{div} p \in L^2(\Omega)\}$.

An explicit solution can be constructed in the case of Dirichlet boundary conditions.

Example 10.4 Let $r > 0$ be such that $B_r(0) \subset \Omega$ and define $g = \chi_{B_r(0)}$. Then

$$u = \max\big\{0, 1 - d/(\alpha r)\big\}\chi_{B_r}(0)$$

is the minimizer for I subject to $u|_{\partial\Omega} = 0$.

Proof Assume that $d \le \alpha r$ and define

$$p(x) = \begin{cases} -r^{-1}x & \text{for} \quad |x| \le r, \\ -rx/|x|^2 & \text{for} \quad |x| \ge r. \end{cases}$$

Then $p \in H(\mathrm{div};\Omega)$ with $\mathrm{div}\ p = -(d/r)\chi_{B_r(0)}$ and $|p| \le 1$. Moreover, we have $u = (1/\alpha)\ \mathrm{div}\ p + g$. Since $p = -n$ on $\partial B_r(0)$ we have for every $q \in H(\mathrm{div};\Omega)$ with $|q| \le 1$ that

$$-(u, \mathrm{div}(q-p)) = -\big(1 - d/(\alpha r)\big)\int_{\partial B_r(0)} (q-p)\cdot n \,\mathrm{d}s \le 0.$$

If $d \ge \alpha r$, we define

$$p(x) = \begin{cases} -(\alpha/d)x & \text{for} \quad |x| \le r, \\ -(\alpha/d)r^2 x/|x|^2 & \text{for} \quad |x| \ge r \end{cases}$$

and verify $\mathrm{div}\ p = -\alpha\chi_{B_r(0)} = -\alpha g$, i.e., $u = (1/\alpha)\ \mathrm{div}\ p + g = 0$, and $|p| \le \alpha r/d \le 1$. Since $u = 0$ the variational inclusion $Du \in \partial I_{K_1(0)}(p)$ is satisfied. □

10.2 Numerical Approximation

We discuss in this section the numerical approximation and iterative solution of the minimization problem defined through the functional I, which for every function $u \in BV(\Omega) \cap L^2(\Omega)$ is given by

$$I(u) = |Du|(\Omega) + \frac{\alpha}{2}\|u - g\|^2$$

for $\alpha > 0$ and $g \in L^2(\Omega)$. The subsequent discussion is based on results in [6–8, 10].

10.2.1 $W^{1,1}$ Conforming Approximation

The finite element space $\mathscr{S}^1(\mathscr{T}_h)$ defines a subspace of $BV(\Omega) \cap W^{1,1}(\Omega)$. Due to the density of smooth functions in $BV(\Omega)$ with respect to intermediate convergence, we can approximate functions in $BV(\Omega)$ by functions in $\mathcal{S}^1(\mathcal{T}_h)$. The following lemma provides bounds on the approximation error. For ease of presentation we restrict to the case $d = 2$.

Lemma 10.1 (Approximation of BV functions) *Assume that $\Omega \subset \mathbb{R}^2$ is star-shaped and let $\varepsilon > 0$. For every $u \in BV(\Omega)$ there exists $u_{\varepsilon,h} \in \mathcal{S}^1(\mathcal{T}_h)$ such that*

$$\|\nabla u_{\varepsilon,h}\|_{L^1(\Omega)} \le (1 + ch\varepsilon^{-1} + c\varepsilon)|Du|(\Omega),$$

and

$$\|u_{\varepsilon,h} - u\|_{L^1(\Omega)} \le c(h^2\varepsilon^{-1} + \varepsilon)|Du|(\Omega).$$

If $u \in L^\infty(\Omega)$, then $\|u_{\varepsilon,h}\|_{L^\infty(\Omega)} \le \|u\|_{L^\infty(\Omega)}$.

Proof Since $C^\infty(\overline{\Omega})$ is dense in $BV(\Omega)$ with respect to intermediate convergence we may choose a function $\widetilde{u} \in C^1(\overline{\Omega})$, such that $\|\widetilde{u} - u\|_{L^1(\Omega)} \le c\varepsilon|Du|(\Omega)$ and $\|\nabla\widetilde{u}\|_{L^1(\Omega)} \le (1 + \varepsilon)|Du|(\Omega)$. Moreover, if $u \in L^\infty(\Omega)$, then we have that $\|\widetilde{u}\|_{L^\infty(\Omega)} \le \|u\|_{L^\infty(\Omega)}$. This allows us to assume $u \in C^1(\overline{\Omega})$ in the following. We suppose that Ω is star-shaped with respect to 0 and define the set $\widehat{\Omega}_\varepsilon = (1 + \varepsilon)\Omega$ and the linear transformation $\phi : \widehat{\Omega}_\varepsilon \to \Omega$, $\widehat{x} \mapsto \widehat{x}/(1 + \varepsilon)$. We set $\widehat{u}_\varepsilon = u \circ \phi$. and with a nonnegative convolution kernel $\rho_\varepsilon \in C^\infty(\mathbb{R}^2)$, we let $u_\varepsilon = (\widehat{u}_\varepsilon * \rho_\varepsilon)|_\Omega$ and define $u_{\varepsilon,h} = \mathcal{I}_h u_\varepsilon$. To prove the estimates we first note that nodal interpolation estimates guarantee

$$\|u_{\varepsilon,h} - u_\varepsilon\|_{L^1(\Omega)} + h\|\nabla(u_{\varepsilon,h} - u_\varepsilon)\| \le ch^2\|D^2u_\varepsilon\|_{L^1(\Omega)}.$$

Standard mollification arguments show that

$$\begin{aligned} \|u_\varepsilon - \widehat{u}_\varepsilon\|_{L^1(\Omega)} &\le c\varepsilon\|\nabla\widehat{u}_\varepsilon\|_{L^1(\widehat{\Omega}_\varepsilon)}, \\ \varepsilon\|D^2u_\varepsilon\|_{L^1(\Omega)} + \|\nabla u_\varepsilon\|_{L^1(\Omega)} &\le \|\nabla\widehat{u}_\varepsilon\|_{L^1(\widehat{\Omega}_\varepsilon)}. \end{aligned}$$

A transformation argument and a direct calculation imply that

$$\begin{aligned} \|\widehat{u}_\varepsilon - u\|_{L^1(\Omega)} &\le c\varepsilon\|\nabla u\|_{L^1(\Omega)}, \\ \|\nabla\widehat{u}_\varepsilon\|_{L^1(\Omega)} &\le (1 + \varepsilon)\|\nabla u\|_{L^1(\Omega)}. \end{aligned}$$

The combination of the estimates proves the asserted bounds for the case $u \in C^1(\overline{\Omega})$. The estimate $\|u_{\varepsilon,h}\|_{L^\infty(\Omega)} \le \|u\|_{L^\infty(\Omega)}$ is a direct consequence of the construction. □

Remarks 10.9 (i) For $d \ge 3$ the same result can be proved by employing a quasi-interpolation operator instead of the nodal interpolation operator.
(ii) The estimate of the lemma and Hölder's inequality imply that for functions $u \in BV(\Omega) \cap L^\infty(\Omega)$ we have $\inf_{v_h \in \mathcal{S}^1(\mathcal{T}_h)} \|u - v_h\|_{L^p(\Omega)} \le ch^{1/p}$ for $1 \le p < \infty$.

(iii) Optimizing the convergence rates of the estimates in the lemma simultaneously for intermediate convergence leads to the choice $\varepsilon = h^{1/2}$ and the suboptimal estimate $\|u - u_{\varepsilon,h}\|_{L^1(\Omega)} \leq ch^{1/2}$.

Since the functional I is strongly convex, the distance of any function to the minimum is controlled by the difference of the values of the functional.

Lemma 10.2 (Convexity) *If $u \in BV(\Omega) \cap L^2(\Omega)$ is the minimizer for I, then we have*

$$\frac{\alpha}{2}\|u - v\|^2 \leq I(v) - I(u)$$

for every $v \in BV(\Omega) \cap L^2(\Omega)$.

Proof We define $F : BV(\Omega) \to \mathbb{R}$ and $G : L^2(\Omega) \to \mathbb{R}$ by

$$F(u) = |Du|(\Omega), \quad G(u) = \frac{\alpha}{2}\|u - g\|^2$$

and extend F by $+\infty$ to $L^2(\Omega)$. Then F is convex and G is strongly convex and Fréchet differentiable with $\delta G(u)[w] = \alpha(u - g, w)$, i.e., we have

$$\delta G(u)[v - u] + \frac{\alpha}{2}\|u - v\|^2 + G(u) = G(v)$$

for all $u, v \in L^2(\Omega)$. Since $u \in BV(\Omega) \cap L^2(\Omega)$ is a minimizer, we have

$$0 \in \partial I(u) = \partial F(u) + \delta G(u),$$

or equivalently $-\delta G(u) \in \partial F(u)$, i.e.,

$$-\delta G(u)[v - u] + F(u) \leq F(v).$$

The strong convexity of G yields

$$\frac{\alpha}{2}\|u - v\|^2 + G(u) - G(v) + F(u) \leq F(v)$$

which proves the assertion. □

Theorem 10.7 (Error estimate) *Assume that $\Omega \subset \mathbb{R}^2$ is star-shaped and $g \in L^\infty(\Omega)$. Let $u \in BV(\Omega) \cap L^2(\Omega)$ and $u_h \in \mathscr{S}^1(\mathscr{T}_h)$ be the minimizers for I in the respective spaces. We then have*

$$\frac{\alpha}{2}\|u - u_h\|^2 \leq ch^{1/2}.$$

Proof Lemma 10.2 and the fact that $I(u_h) \leq I(v_h)$ for all $v_h \in \mathscr{S}^1(\mathscr{T}_h)$ imply that

$$\begin{aligned}\frac{\alpha}{2}\|u-u_h\|^2 &\le I(u_h)-I(u)\le I(v_h)-I(u)\\ &= \|\nabla v_h\|_{L^1(\Omega)}-|Du|(\Omega)\\ &\quad+\frac{\alpha}{2}\int\limits_{\Omega}\big((v_h-g)-(u-g)\big)\big((v_h-g)+(u-g)\big)\,\mathrm{d}x\\ &\le \|\nabla v_h\|_{L^1(\Omega)}-|Du|(\Omega)+\frac{\alpha}{2}\|v_h-u\|_{L^1(\Omega)}\|v_h+u-2g\|_{L^\infty(\Omega)}.\end{aligned}$$

For $\varepsilon>0$ we let $v_h=u_{\varepsilon,h}\in\mathscr{S}^1(\mathscr{T}_h)$ be an approximation of u as in Lemma 10.1 and deduce that

$$\frac{\alpha}{2}\|u-u_h\|^2\le c(h\varepsilon^{-1}+\varepsilon)|Du|(\Omega)+c(h^2\varepsilon^{-1}+\varepsilon)|Du|(\Omega).$$

With $\varepsilon=h^{1/2}$ we find the asserted bound. □

Remarks 10.10 (i) Since for $u\in BV(\Omega)\cap L^2(\Omega)$ the best approximation in $\mathscr{S}^1(\mathscr{T}_h)$ satisfies $\inf_{v_h\in\mathscr{S}^1(\mathscr{T}_h)}\|u-v_h\|\le h^{1/2}$, the convergence rate of the theorem is suboptimal. Numerical experiments indicate that the optimal convergence rate $\mathscr{O}(h^{1/2})$ in $L^2(\Omega)$ is in general not attained.
(ii) If $\Omega=(a,b)\subset\mathbb{R}$ and the minimizer $u\in BV(\Omega)\cap L^2(\Omega)$ is piecewise continuous, then we can employ the nodal interpolant $v_h=\mathscr{I}_h u$ in the proof of the theorem and noting that $\|\nabla\mathscr{I}_h u\|\le|Du|(\Omega)$ and $\|u-\mathscr{I}_h u\|_{L^1(\Omega)}\le ch|Du|(\Omega)$, we obtain the quasi-optimal estimate $\|u-u_h\|\le ch^{1/2}$.

The best approximation result $\inf_{v_h\in\mathscr{S}^1(\mathscr{T}_h)}\|u-v_h\|_{L^p(\Omega)}\le ch^{1/p}$ for functions $u\in BV(\Omega)\cap L^\infty(\Omega)$ can in general not be improved as the following example shows.

Example 10.5 Let $\Omega=(-1,1)$, $\mathscr{T}_h$ a uniform triangulation of Ω with mesh-size $h>0$ such that $z=0$ is a node of $\mathscr{T}_h$. For $u=\mathrm{sign}$, we then have

$$\inf_{v_h\in\mathscr{S}^1(\mathscr{T}_h)}\|u-v_h\|_{L^p(\Omega)}\ge ch^{1/p}.$$

To prove this we show that the entire approximation error is concentrated at the discontinuity at $x=0$. We assume that there exists a minimal $w_h\in\mathscr{S}^1(\mathscr{T}_h)$ which is antisymmetric, i.e., we have $w_h(-x)=-w_h(x)$ for $x\in(0,1)$ and $w_h(0)=0$. Then the function w_h is affine on $(-h,h)$ with slope $a/h\in\mathbb{R}$, cf. Fig. 10.3, for some $a\in\mathbb{R}$, and we have with the transformation $y=x/h$ that

$$\int\limits_{(-h,h)}|u-w_h|^p\,\mathrm{d}x=2\int\limits_{(0,h)}|1-ax/h|^p\,\mathrm{d}x=2h\int\limits_{(0,1)}|1-ay|^p\,\mathrm{d}y.$$

The value of the integral related to the minimizing choice of a is positive and independent of h which implies that $\|u-w_h\|_{L^p(\Omega)}\ge ch^{1/p}$.

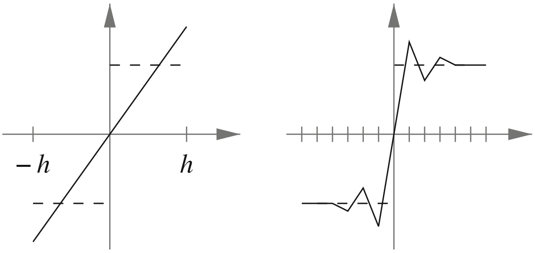


Fig. 10.3 The approximation of a discontinuous function with continuous, piecewise affine functions leads to an error $\|u - w_h\|_{L^p(\Omega)} \geq ch^{1/p}$ (*left*); for the best approximation of $u = \text{sign}$ in $\mathscr{S}^1(\mathscr{T}_h)$ with respect to the L^2 norm, the Gibb's phenomenon occurs at the discontinuity (*right*)

10.2.2 Piecewise Constant Approximation

The set of piecewise constant finite element functions $\mathscr{L}^0(\mathscr{T}_h)$ is a subset of $BV(\Omega)$. It is straightforward to check that for a sequence of triangulations their union defines a dense subset with respect to weak convergence. We will show that density with respect to intermediate convergence fails and hence that the discretization of the model problem with piecewise constant finite elements may not approximate the right minimum.

Proposition 10.5 (Piecewise constant functions) *For every $u_h \in \mathscr{L}^0(\mathscr{T}_h)$ we have*

$$|Du_h|(\Omega) = \sum_{S \in \mathscr{S}_h \cap \Omega} \|[\![u_h]\!]\|_{L^1(S)}.$$

Proof The identity follows directly from an elementwise integration by parts. □

Proposition 10.6 (Nonapproximation) *Let $\Omega = (-1/2, 1/2) \times (0, 1)$ and let $u \in BV(\Omega) \cap L^\infty(\Omega)$ be, for $x = (x_1, x_2) \in \Omega$, defined by $u(x_1, x_2) = \chi_{\{x_1 < 0\}}$. For each $n \geq 1$ let $\mathscr{T}_n$ be the triangulation of Ω with maximal mesh-size $h_n = 1/n$, as shown in Fig. 10.4. Then there is no sequence $(u_n)_{n \in \mathbb{N}} \subset L^1(\Omega)$ with $u_n \in \mathscr{L}^0(\mathscr{T}_n)$ for all $n \in \mathbb{N}$ such that $u_n \to u$ in $L^1(\Omega)$ and $|Du_n|(\Omega) \to |Du|(\Omega) = 1$ as $n \to \infty$.*

Proof Let $(u_n)_{n \in \mathbb{N}}$ be a sequence with $u_n \in \mathscr{L}^0(\mathscr{T}_n)$ such that $\|u_n - u\|_{L^1(\Omega)} \to 0$ and $|Du_n|(\Omega) \leq c$ for all $n \in \mathbb{N}$. Given $n \in \mathbb{N}$ we define the sets R_j^n for $j = 1, 2, \ldots, n$ by

$$R_j^n = \{(x_1, x_2) \in \Omega : (j-1)/n < x_2 < j/n\}$$

and set $R^n = R_1^n$. Let $\overline{u}_n \in L^1(R^n)$ be the average of u_n over all strips, i.e., for $(x_1, x_2) \in R^n$ set

$$\overline{u}_n(x_1, x_2) = \frac{1}{n} \sum_{j=1}^{n} u_n(x_1, x_2 + j/n),$$

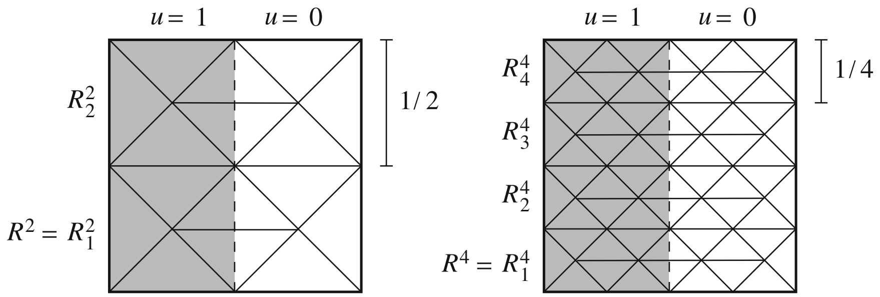


Fig. 10.4 Construction of triangulations $\mathscr{T}_n$, $n \in \mathbb{N}$, of $\Omega = (-1/2, 1/2) \times (0, 1)$ on which piecewise constant finite element functions are not dense in $BV(\Omega)$ with respect to intermediate convergence; the jump set of the function $u = \chi_{\{x_1 < 0\}}$ is not resolved by the triangulations

and reflect $\overline{u}_n$ across the x_1-axis, i.e., $\overline{u}_n(x_1, -x_2) = \overline{u}_n(x_1, x_2)$ for $(x_1, x_2) \in R^n$. We then define $\widetilde{u}_n \in L^1(\Omega)$ by periodically extending $\overline{u}_n$ with period $2/n$ in the x_2-direction. Then $\widetilde{u}_n \in L^1(\Omega)$ is continuous across the interfaces $\overline{R}_j^n \cap \overline{R}_{j+1}^n$ for $j = 1, 2, \ldots, n-1$ and we have $\|\widetilde{u}_n - u\|_{L^1(R_j^n)} = \|\overline{u}_n - u\|_{L^1(R^n)}$ and $|D\widetilde{u}_n|(R_j^n) = |D\overline{u}_n|(R^n)$ for $j = 1, 2, \ldots, n$, where $|D\overline{u}_n|(R^n)$ denotes the total variation of $D\overline{u}_n$ on R^n. With the triangle inequality we verify that

$$
\begin{aligned}
|D\widetilde{u}_n|(\Omega) &= n|D\overline{u}_n|(R^n) \le |Du_n|(\Omega), \\
\|\widetilde{u}_n - u\|_{L^1(\Omega)} &= n\|\overline{u}_n - u\|_{L^1(R^n)} \le \|u_n - u\|_{L^1(\Omega)}.
\end{aligned}
$$

For every $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $\|u_n - u\|_{L^1(\Omega)} < \varepsilon$ for all $n \ge N$, i.e.,

$$
\|\overline{u}_n - u\|_{L^1(R^n)} < \varepsilon/n.
$$

For each $n \ge N$ there exist distinct triangles $T_+^1, T_+^2, T_-^1, T_-^2 \in \mathscr{T}_n \cap R^n$ with $\overline{u}_n|_{T_+^1 \cup T_+^2} \ge 1 - 4\varepsilon$ and $\overline{u}_n|_{T_-^1 \cup T_-^2} \le 4\varepsilon$ since otherwise $\|\overline{u}_n - u\|_{L^1(R^n)} \ge \varepsilon/n$. The triangle inequality along disjoint paths of neighboring elements connecting T_-^j and T_+^j for $j = 1, 2$, respectively, yields that

$$
\begin{aligned}
(1 - 8\varepsilon)\sqrt{2}/n &\le (h_n/\sqrt{2})\Big(\big|\overline{u}_n|_{T_-^1} - \overline{u}_n|_{T_+^1}\big| + \big|\overline{u}_n|_{T_-^2} - \overline{u}_n|_{T_+^2}\big|\Big) \\
&\le \sum_{S \in \mathscr{S}_h \cap R^n} \|[\![\overline{u}_n]\!]\|_{L^1(S)} = |D\overline{u}_n|(R^n)
\end{aligned}
$$

and hence $|Du_n|(\Omega) \ge |D\widetilde{u}_n|(\Omega) \ge (1 - 8\varepsilon)\sqrt{2}$ for all $n \ge N$, i.e., we have that $|Du_n|(\Omega) \not\to 1 = |Du|(\Omega)$ as $n \to \infty$. □

10.2.3 Iterative Solution

To develop an iterative solution method for the nondifferentiable minimization problem, we first state optimality conditions for the minimization of I in $\mathscr{S}^1(\mathscr{T}_h)$. For this we note that the minimization of I can be equivalently expressed as a saddle-point problem; that is, due to the fact that ∇u_h is elementwise constant for $u_h \in \mathscr{S}^1(\mathscr{T}_h)$ we have

$$\begin{aligned}\inf_{u_h \in \mathscr{S}^1(\mathscr{T}_h)} \int_\Omega |\nabla u_h| \,\mathrm{d}x + \frac{\alpha}{2}\|u_h - g\|^2 &= \inf_{u_h \in \mathscr{S}^1(\mathscr{T}_h)} \sup_{p_h \in \mathscr{L}^0(\mathscr{T}_h)^d} \int_\Omega p_h \cdot \nabla u_h \,\mathrm{d}x \\ &\quad + \frac{\alpha}{2}\|u_h - g\|^2 - I_{K_1(0)}(p_h) \\ &= \inf_{u_h \in \mathscr{S}^1(\mathscr{T}_h)} \sup_{p_h \in \mathscr{L}^0(\mathscr{T}_h)^d} L_h(u_h, p_h),\end{aligned}$$

where $I_{K_1(0)}$ is the indicator functional of the set $K_1(0) = \{p \in L^\infty(\Omega; \mathbb{R}^d) : |p| \le 1 \text{ a.e. in } \Omega\}$.

Lemma 10.3 (Optimality) *The function $u_h \in \mathscr{S}^1(\mathscr{T}_h)$ minimizes I in $\mathscr{S}^1(\mathscr{T}_h)$ if and only if there exists $p_h \in \mathscr{L}^0(\mathscr{T}_h)^d$ with $|p_h| \le 1$ in Ω such that*

$$(p_h, \nabla v_h) = -\alpha(u_h - g, v_h), \quad (\nabla u_h, q_h - p_h) \le 0$$

for all $(v_h, q_h) \in \mathscr{S}^1(\mathscr{T}_h) \times \mathscr{L}^0(\mathscr{T}_h)^d$ with $|q_h| \le 1$ in Ω.

Proof The existence of a saddle point $(u_h, p_h) \in \mathscr{S}^1(\mathscr{T}_h) \times \mathscr{L}^0(\mathscr{T}_h)^d$ follows from the fact that the Lagrangian function L_h is a lower-semicontinuous, proper, convex-concave function, cf., e.g., [14] for details. The equations are the corresponding Kuhn–Tucker optimality conditions, i.e.,

$$0 = \delta_{u_h} L_h(u_h, p_h), \quad 0 \in \partial_{p_h} L_h(u_h, p_h),$$

where we note that $\xi_h \in \partial I_{K_1(0)}(p_h)$ for $\xi_h \in \mathscr{L}^0(\mathscr{T}_h)^d$ and $p_h \in \mathscr{L}^0(\mathscr{T}_h)^d \cap K_1(0)$, i.e.,

$$(\xi_h, q_h - p_h) + I_{K_1(0)}(p_h) \le I_{K_1(0)}(q_h)$$

for all $q_h \in \mathscr{L}^0(\mathscr{T}_h)^d$, if and only if

$$(\xi_h, q_h - p_h) \le 0$$

for all $q_h \in \mathscr{L}^0(\mathscr{T}_h)^d \cap K_1(0)$. □

To find a saddle point for L_h we use a descent flow with respect to u_h and an ascent flow with respect to p_h, i.e.,

$$\partial_t u_h = -\delta_{u_h} L_h(u_h, p_h), \quad \partial_t p_h \in \partial_{p_h} L_h(u_h, p_h).$$

With an appropriate time-discretization and a discrete inner product $(\cdot,\cdot)_{h,s}$ on $\mathscr{S}^1(\mathscr{T}_h)$ that may differ from the L^2 inner product, this motivates the following iteration which specifies the abstract primal-dual iteration of Algorithm 4.5.

Algorithm 10.1 (*Primal-dual iteration*) Let $(\cdot,\cdot)_{h,s}$ be an inner product on $\mathscr{S}^1(\mathscr{T}_h)$, $\tau > 0$, $(u_h^0, p_h^0) \in \mathscr{S}^1(\mathscr{T}_h) \times \mathscr{L}^0(\mathscr{T}_h)^d$, set $d_t u_h^0 = 0$, and for $k = 0, 1, \dots$ with $\widetilde{u}_h^k = u_h^{k-1} + \tau d_t u_h^{k-1}$ solve the equations

$$\begin{aligned}(-d_t p_h^k + \nabla \widetilde{u}_h^k, q_h - p_h^k) &\le 0,\\ (d_t u_h^k, v_h)_{h,s} + (p_h^k, \nabla v_h) + \alpha(u_h^k - g, v_h) &= 0\end{aligned}$$

subject to $|p_h^k| \le 1$ in Ω for all $(v_h, q_h) \in \mathscr{S}^1(\mathscr{T}_h) \times \mathscr{L}^0(\mathscr{T}_h)^d$ with $|q_h| \le 1$ in Ω. Stop the iteration if $\|d_t u_h^k\|_{h,s} \le \varepsilon_{\mathrm{stop}}$.

Remark 10.11 Notice that p_h^k is the unique minimizer of the mapping

$$q_h \mapsto \frac{1}{2\tau}\|q_h - p_h^{k-1}\|^2 - (q_h, \nabla \widetilde{u}_h^k) + I_{K_1(0)}(q_h)$$

and given by the truncation operation

$$p_h^k = \big(p_h^{k-1} + \tau \nabla \widetilde{u}_h^k\big) / \max\{1, |p_h^{k-1} + \tau \nabla \widetilde{u}_h^k|\}$$

which can be computed elementwise.

The iterates of Algorithm 10.1 converge to a stationary point if τ is sufficiently small.

Proposition 10.7 (Convergence) *Let $u_h \in \mathscr{S}^1(\mathscr{T}_h)$ be minimal for I in $\mathscr{S}^1(\mathscr{T}_h)$ and define*

$$\theta = \sup_{v_h \in \mathscr{S}^1(\mathscr{T}_h)\setminus\{0\}} \frac{\|\nabla v_h\|}{\|v_h\|_{h,s}}.$$

If $\tau\theta \le 1$, then the iterates of Algorithm 10.1 converge to u_h in the sense that they satisfy for every $N \ge 1$

$$\tau \sum_{k=1}^N \Big((1-\tau^2\theta^2)\frac{\tau}{2}\|d_t u_h^k\|_{h,s}^2 + \alpha\|u_h - u_h^k\|^2\Big) \le \frac{1}{2}\big(\|u_h - u_h^0\|_{h,s}^2 + \|p_h - p_h^0\|^2\big).$$

Proof Let $p_h \in \mathscr{L}^0(\mathscr{T}_h)^d$ be as in Lemma 10.3. Upon choosing $v_h = u_h - u_h^k$ and $q_h = p_h$ in Algorithm 10.1, we find that

$$\begin{aligned}&\frac{d_t}{2}\big(\|u_h - u_h^k\|_{h,s}^2 + \|p_h - p_h^k\|^2\big) + \frac{\tau}{2}\big(\|d_t u_h^k\|_{h,s}^2 + \|d_t p_h^k\|^2\big) + \alpha\|u_h - u_h^k\|^2\\ &\quad = -(d_t u_h^k, u_h - u_h^k)_{h,s} - (d_t p_h^k, p_h - p_h^k) + \alpha\|u_h - u_h^k\|^2\end{aligned}$$

$$\le (p_h^k, \nabla(u_h - u_h^k)) + \alpha(u_h^k - g, u_h - u_h^k) - (p_h - p_h^k, \nabla\widetilde{u}_h^k) + \alpha\|u_h - u_h^k\|^2.$$

Using that

$$(u_h^k - g, u_h - u_h^k) + \|u_h - u_h^k\|^2 = (u_h - g, u_h - u_h^k)$$

and choosing $q_h = p_h^k$ in Lemma 10.3, we deduce that

$$\begin{aligned}
&\frac{d_t}{2}\big(\|u_h - u_h^k\|_{h,s}^2 + \|p_h - p_h^k\|^2\big) + \frac{\tau}{2}\big(\|d_t u_h^k\|_{h,s}^2 + \|d_t p_h^k\|^2\big) + \alpha\|u_h - u_h^k\|^2 \\
&\quad = (p_h^k, \nabla(u_h - u_h^k)) - (p_h - p_h^k, \nabla\widetilde{u}_h^k) + \alpha(u_h - g, u_h - u_h^k) \\
&\quad = (p_h^k, \nabla(u_h - u_h^k)) - (p_h - p_h^k, \nabla\widetilde{u}_h^k) - (p_h, \nabla(u_h - u_h^k)) \\
&\quad = (p_h - p_h^k, \nabla(u_h^k - \widetilde{u}_h^k)) + (p_h^k - p_h, \nabla u_h) \\
&\quad \le (p_h - p_h^k, \nabla(u_h^k - \widetilde{u}_h^k)) = \tau^2(p_h - p_h^k, \nabla d_t^2 u_h^k),
\end{aligned}$$

where we used $u_h^k - \widetilde{u}_h^k = \tau^2 d_t^2 u_h^k$ in the last identity. Multiplication by τ, summation over $k = 1, 2, \ldots, K$, discrete integration by parts, Young's inequality, and $d_t u_h^0 = 0$ show that for the right-hand side we have

$$\begin{aligned}
\tau^3 \sum_{k=1}^K (p_h - p_h^k, \nabla d_t^2 u_h^k) &= \tau^3 \sum_{k=1}^K (d_t p_h^k, \nabla d_t u_h^{k-1}) + \tau^2 (p_h - p_h^k, \nabla d_t u_h^k)\big|_{k=0}^K \\
&\le \frac{\tau^2}{2}\Big(\sum_{k=1}^K \tau^2 \|\nabla d_t u_h^{k-1}\|^2 + \|d_t p_h^k\|^2\Big) \\
&\quad + \frac{1}{2}\|p_h - p_h^K\|^2 + \frac{\tau^4}{2}\|\nabla d_t u_h^K\|^2 \\
&\le \frac{\tau^2}{2}\Big(\sum_{k=1}^K \tau^2\theta^2 \|d_t u_h^{k-1}\|_{h,s}^2 + \|d_t p_h^k\|^2\Big) \\
&\quad + \frac{1}{2}\|p_h - p_h^K\|^2 + \frac{\tau^4\theta^2}{2}\|d_t u_h^K\|_{h,s}^2.
\end{aligned}$$

Due to the assumption $\tau\theta \le 1$ we may absorb the terms of the right-hand side and conclude that

$$\begin{aligned}
&\frac{1}{2}\|u_h - u_h^K\|_{h,s}^2 + \tau\sum_{k=1}^K \frac{\tau}{2}(1 - \tau\theta^2)\|d_t u_h^k\|^2 + \tau\sum_{k=1}^K \alpha\|u_h - u_h^k\|^2 \\
&\quad \le \frac{1}{2}\big(\|u_h - u_h^0\|_{h,s}^2 + \|p_h - p_h^0\|^2\big).
\end{aligned}$$

This proves the theorem. □

Remark 10.12 Notice that we cannot expect convergence $p_h^n \to p_h$ since p_h is not unique in general, e.g., if $\nabla u_h|_T = 0$ for some $T \in \mathscr{T}_h$.

Useful choices of the inner product $(\cdot,\cdot)_{h,s}$ are weighted combinations of the inner product in $L^2(\Omega)$ and the semi-inner product in $H^1(\Omega)$.

Proposition 10.8 (Discrete inner products) *For* $s \in [0,1]$ *and* $v_h, w_h \in \mathscr{S}^1(\mathscr{T}_h)$ *define*

$$(v_h, w_h)_{h,s} = (v_h, w_h) + h_{\min}^{(1-s)/s}(\nabla v_h, \nabla w_h),$$

where $h_{\min}^{(1-s)/s} = 0$ *for* $s = 0$. *We then have* $\|\nabla v_h\| \le c h_{\min}^{-\min\{1,(1-s)/(2s)\}} \|v_h\|_{h,s}$ *for all* $v_h \in \mathscr{S}^1(\mathscr{T}_h)$ *with* $c = 1$ *if* $s > 0$.

Proof If $s > 0$, then we have by definition of $\|v_h\|_{h,s}^2 = (v_h, v_h)_{h,s}$ that

$$\|\nabla v_h\|^2 \le h_{\min}^{-(1-s)/s} \|v_h\|_{h,s}^2$$

for all $v_h \in \mathscr{S}^1(\mathscr{T}_h)$. For $s \ge 0$ the inverse estimate $\|\nabla v_h\| \le c h_{\min}^{-1} \|v_h\|$, valid for all $v_h \in \mathscr{S}^1(\mathscr{T}_h)$, implies the assertion. □

To fully justify the choice of the scalar products $(\cdot,\cdot)_{h,s}$ for $s > 0$, we have to show that the right-hand side in the estimate of Proposition 10.7 is bounded h-independently. For $s \le 1/2$ this is guaranteed by the following lemma if the sequence $(u_h)_{h>0}$ of finite element approximations is uniformly bounded in the set $W^{1,1}(\Omega) \cap L^\infty(\Omega)$.

Lemma 10.4 (Discrete interpolation estimate) *For every* $v_h \in \mathscr{S}^1(\mathscr{T}_h)$ *we have*

$$h_{\min} \|\nabla v_h\|_{L^2(\Omega)}^2 \le c \|v_h\|_{L^\infty(\Omega)} \|\nabla v_h\|_{L^1(\Omega)}.$$

Proof For $T \in \mathscr{T}_h$, an integration by parts on T together with the fact that $\Delta v_h|_T = 0$, implies that

$$h_T \int_T |\nabla v_h|^2 \,\mathrm{d}x = h_T \int_{\partial T} v_h \nabla v_h \cdot n_T \,\mathrm{d}s \le h_T |\partial T| \|v_h\|_{L^\infty(T)} |T|^{-1} \|\nabla v_h\|_{L^1(T)}.$$

Noting $h_T|\partial T| \le c|T|$, a summation over $T \in \mathscr{T}_h$ implies the assertion. □

Remark 10.13 To obtain approximations with residuals that are bounded independently of the parameter s, the stopping criterion

$$\sup_{v_h \in \mathscr{S}^1(\mathscr{T}_h)} \frac{(d_t u_h^k, v_h)_{h,s}}{\|v_h\|} \le \varepsilon_{\text{stop}}$$

should be used.

10.2.4 Realization

The MATLAB code displayed in Fig. 10.5 is an implementation of the primal dual method of Algorithm 10.1 with the scalar product $(\cdot,\cdot)_{h,1/2}$ defined in Proposition 10.8 and the corresponding choice $\tau = h^{1/2}/10$. It computes the update of p_h^{k-1} via the elementwise operation

$$p_h^k = \frac{p_h^{k-1} + \tau \nabla \widetilde{u}_h^{k-1}}{\max\{1, |p_h^{k-1} + \tau \nabla \widetilde{u}_h^{k-1}|\}}$$

and the linear system of equations

$$(d_t u_h^k, v_h)_{h,s} + (p_h^k, \nabla v_h) = -\alpha(u_h^k - g, v_h)$$

for all $v_h \in \mathscr{S}^1(\mathscr{T}_h)$. The second term on the left-hand side is represented by the matrix with the entries

$$(\chi_T e^\ell, \nabla \varphi_z) = |T|\, \partial_\ell \varphi_z|_T$$

for all $T \in \mathscr{T}_h$, $\ell = 1, 2, \ldots, d$, and $z \in \mathscr{N}_h$ which is assembled in the routine `mixed_matrix`.

10.2.5 A Posteriori Error Control

We apply the abstract framework for a posteriori error estimates for strongly convex minimization problems of Theorem 4.2 to control the approximation error in the numerical minimization of I. The estimate states that the distance of an arbitrary approximation to the minimizer is controlled by the primal-dual gap. The dual functional is for $p \in H_N(\mathrm{div};\Omega)$ given by

$$D(p) = -\frac{1}{2\alpha}\| \operatorname{div} p + \alpha g\|^2 + \frac{\alpha}{2}\|g\|^2 - I_{K_1(0)}(p),$$

and we have $D(q) \le I(u)$ for every $q \in H_N(\mathrm{div};\Omega)$ with equality for a solution of the dual problem.

Theorem 10.8 (A posteriori error estimate) *Let $u \in BV(\Omega) \cap L^2(\Omega)$ be the minimizer for I. Then for every $u_h \in \mathscr{S}^1(\mathscr{T}_h)$ and $\widehat{p}_h \in H_N(\mathrm{div};\Omega)$ with $|\widehat{p}_h| \le 1$, we have*

$$\frac{\alpha}{2}\|u - u_h\|^2 \le \|\nabla u_h\|_{L^1(\Omega)} - \int_\Omega \nabla u_h \cdot \widehat{p}_h \,\mathrm{d}x + \frac{1}{2\alpha}\| \operatorname{div} \widehat{p}_h - \alpha(u_h - g)\|^2.$$

```
function tv_reg_primal_dual(d,red)
[c4n,n4e,Db,Nb] = triang_cube(d); c4n = c4n-.5;
for j = 1:red
    [c4n,n4e,Db,Nb] = red_refine(c4n,n4e,Db,Nb);
end
h = 2^(-red); alpha = 100; tau = h^(1/2)/10; noise = .4;
[s,m,¬,¬] = fe_matrices(c4n,n4e);
ms = mixed_matrix(c4n,n4e);
A = m+h*s;
[nC,d] = size(c4n); nE = size(n4e,1);
gg = g(c4n)+noise*(rand(nC,1)-.5);
u = zeros(nC,1); u_tilde = u; p = zeros(nE,d);
corr = 1; eps_stop = 1e-2;
while corr > eps_stop
    du_tilde = comp_gradient(c4n,n4e,u_tilde);
    p_tmp = p+tau*du_tilde;
    p = p_tmp./max(1,(sqrt(sum(p_tmp.^2,2))*ones(1,d)));
    P = reshape(p',d*nE,1);
    u_new = (A+tau*alpha*m)\(A*u-tau*ms*P+tau*alpha*m*gg);
    dt_u = (u-u_new)/tau;
    corr = sqrt(dt_u'*A*dt_u)
    u_tilde = 2*u_new-u;
    u = u_new;
    show_p1(c4n,n4e,Db,Nb,u);
end

function ms = mixed_matrix(c4n,n4e)
[nC,d] = size(c4n); nE = size(n4e,1);
ctr = 0; ctr_max = d*(d+1)*nE;
I = zeros(ctr_max,1); J = zeros(ctr_max,1); X = zeros(ctr_max,1);
for j = 1:nE
    X_T = [ones(1,d+1);c4n(n4e(j,:),:)'];
    grads_T = X_T\[zeros(1,d);eye(d)];
    vol_T = det(X_T)/factorial(d);
    for k = 1:d+1
        for ell = 1:d
            ctr = ctr+1;
            I(ctr) = n4e(j,k); J(ctr) = (j-1)*d+ell;
            X(ctr) = vol_T*grads_T(k,ell);
        end
    end
end
ms = sparse(I,J,X,nC,d*nE);

function val = g(x)
val = zeros(size(x,1),1);
val(sqrt(sum(x.^2,2))<.2) = 1;
```

Fig. 10.5 MATLAB realization of Algorithm 10.1 for the iterative minimization of the total variation regularization problem

Proof We recall from Lemma 10.2 that

$$\frac{\alpha}{2}\|u-u_h\|^2 \leq I(u_h)-I(u).$$

Incorporating the duality principle $I(u) \geq D(\widehat{p}_h)$ for all $\widehat{p}_h \in H_N(\mathrm{div};\Omega)$, we deduce that

$$\frac{\alpha}{2}\|u-u_h\|^2 \leq \|\nabla u_h\|_{L^1(\Omega)}+\frac{\alpha}{2}\|u_h-g\|^2+\frac{1}{2\alpha}\|\operatorname{div}\widehat{p}_h+\alpha g\|^2-\frac{\alpha}{2}\|g\|^2+I_{K_1(0)}(\widehat{p}_h).$$

We assume that $|\widehat{p}_h| \leq 1$ in Ω and with straightforward calculations deduce that

$$\begin{aligned}\frac{\alpha}{2}\|u-u_h\|^2 &\leq \|\nabla u_h\|_{L^1(\Omega)}+\frac{1}{2\alpha}\|\operatorname{div}\widehat{p}_h-\alpha(u_h-g)\|^2\\ &\qquad+\int_\Omega u_h(\operatorname{div}\widehat{p}_h+\alpha g)\,\mathrm{d}x+\frac{\alpha}{2}\|u_h-g\|^2-\frac{\alpha}{2}\|g\|^2-\frac{\alpha}{2}\|u_h\|^2\\ &= \|\nabla u_h\|_{L^1(\Omega)}+\frac{1}{2\alpha}\|\operatorname{div}\widehat{p}_h-\alpha(u_h-g)\|^2+\int_\Omega u_h\operatorname{div}\widehat{p}_h\,\mathrm{d}x.\end{aligned}$$

An integration by parts proves the asserted estimate. □

Remarks 10.14 (i) The error estimate is sharp in the sense that if $u=u_h$ and $\widehat{p}_h=p$ solves the dual problem, then the right-hand side vanishes.
(ii) The practical application requires us to compute a conforming approximate solution of the dual problem. The piecewise constant approximation provided by Algorithm 10.1 in general does not satisfy $\widehat{p}_h \in H_N(\mathrm{div};\Omega)$.
(iii) The error estimate gives rise to the nonnegative refinement indicators

$$\eta_T(u_h,\widehat{p}_h)=\|\nabla u_h\|_{L^1(T)}-\int_T \nabla u_h\cdot\widehat{p}_h\,\mathrm{d}x+\frac{1}{2\alpha}\|\operatorname{div}\widehat{p}_h-\alpha(u_h-g)\|^2_{L^2(T)}$$

for $u_h \in \mathscr{S}^1(\mathscr{T}_h)$ and $\widehat{p}_h \in H_N(\mathrm{div};\Omega)$ with $|\widehat{p}_h| \leq 1$. Noting the optimality condition $\operatorname{div} p=\alpha(u-g)$ and the duality relation

$$|Du|(\Omega)=-\int_\Omega u\operatorname{div} p\,\mathrm{d}x$$

for an exact solution $(u,p)\in\big(BV(\Omega)\cap L^2(\Omega)\big)\times H_N(\mathrm{div};\Omega)$ with $|p|\leq 1$ in Ω, the refinement indicators have the interpretation of a residual.

10.2.6 Regularized Minimization

In some situations a regularized treatment of the functional I provides accurate approximations and in this case a semi-implicit discretization of the corresponding gradient flow defines a useful iterative scheme. We define the regularized functional I_δ for $\delta > 0$ by

$$I_\delta(u) = \int_\Omega |\nabla u|_\delta \, \mathrm{d}x + \frac{\alpha}{2}\|u-g\|^2$$

for $u \in W^{1,1}(\Omega) \cap L^2(\Omega)$ and with $|p|_\delta = (|p|^2 + \delta^2)^{1/2}$ for every $p \in \mathbb{R}^d$.

Algorithm 10.2 (*Semi-implicit, regularized L^2-flow*) Given $\delta > 0$, $\tau > 0$, and $u_h^0 \in \mathscr{S}^1(\mathscr{T}_h)$ compute the sequence $(u_h^k)_{k=0,1,\dots}$ by solving

$$(d_t u_h^k, v_h) + \big(|\nabla u_h^{k-1}|_\delta^{-1} \nabla u_h^k, \nabla v_h\big) = -\alpha(u_h^k - g, v_h)$$

for all $v_h \in \mathscr{S}^1(\mathscr{T}_h)$. Stop if $\|d_t u_h^k\| \le \varepsilon_{\text{stop}}$.

Remark 10.15 The choice $v_h = u_h^k$ shows that the iteration is unconditionally weakly stable in the sense that

$$\frac{d_t}{2}\|u_h^k\|^2 + \frac{\tau}{2}\|d_t u_h^k\|^2 + \big\| |\nabla u_h^{k-1}|_\delta^{-1/2} \nabla u_h^k \big\|^2 + \frac{\alpha}{2}\|u_h^k\|^2 \le \frac{\alpha}{2}\|g\|^2$$

for all $k \ge 1$. In order to obtain accurate approximations, the step size should be chosen so that $\tau \le c h_{\min}$. This scaling leads to practically strongly stable approximation schemes for $\delta > 0$ in the sense that the regularized energy I_δ decreases.

If $\delta \le c h^{1/2}$, we have the same error estimates as for the unregularized approximation.

Proposition 10.9 (Regularized approximation) *Let $u \in BV(\Omega) \cap L^2(\Omega)$ be the minimizer for I and let $u_{\delta,h} \in \mathscr{S}^1(\mathscr{T}_h)$ be minimal for*

$$I_\delta(v_h) = \int_\Omega |\nabla v_h|_\delta \, \mathrm{d}x + \frac{\alpha}{2}\|v_h - g\|^2$$

in the set of functions $v_h \in \mathscr{S}^1(\mathscr{T}_h)$. If $\delta \le c h^{1/2}$, then we have

$$\frac{\alpha}{2}\|u - u_{\delta,h}\|^2 \le c h^{1/2}.$$

Proof We first note that for every $p \in \mathbb{R}^d$ we have

$$|p| \le |p|_\delta \le |p| + \delta.$$

With Lemma 10.2 and the fact that $u_{\delta,h}$ is minimal for I_δ in $\mathscr{S}^1(\mathscr{T}_h)$ it follows for every $v_h \in \mathscr{S}^1(\mathscr{T}_h)$ that

$$\begin{aligned}\frac{\alpha}{2}\|u - u_{\delta,h}\|^2 &\le I(u_{\delta,h}) - I(u) \le I_\delta(u_{\delta,h}) - I(u) \le I_\delta(v_h) - I(u)\\ &= I_\delta(v_h) - I(v_h) + I(v_h) - I(u) \le \delta|\Omega| + I(v_h) - I(u).\end{aligned}$$

With $v_h = u_{\varepsilon,h}$, as in Lemma 10.1 for $\varepsilon = h^{1/2}$, we deduce the asserted bound. □

Remark 10.16 An alternative definition for $|p|_\delta$ is given by

$$|p|_\delta = \begin{cases} |p| & \text{if } \ |p| \ge \delta, \\ (|p|^2 + \delta^2)/2 & \text{if } \ |p| \le \delta. \end{cases}$$

Figure 10.6 displays an implementation of Algorithm 10.2. The weighted stiffness matrix is computed in the routine `fe_matrices_weighted` which provides for elementwise constant functions $a, b : \Omega \to \mathbb{R}$ the matrices with entries

$$s_{a,zy} = \int_\Omega a\,\nabla\varphi_z \cdot \nabla\varphi_y \,\mathrm{d}x, \quad m_{b,zy} = \int_\Omega b\,\varphi_z\varphi_y \,\mathrm{d}x$$

for $z, y \in \mathscr{N}_h$.

10.2.7 Total Variation Flow

The total variation arises in various mathematical models describing evolution problems by subdifferential flows. The evolution problems are also often the basis for numerical minimization algorithms. An implicit discretization leads to the following algorithm.

Algorithm 10.3 (*Subdifferential flow*) Given $u_h^0 \in \mathscr{S}^1(\mathscr{T}_h)$ and $\tau > 0$, compute the sequence $(u_h^k)_{k=0,\dots,K} \subset \mathscr{S}^1(\mathscr{T}_h)$ by minimizing for $k = 1, 2, \dots, K$ the functionals

$$I_{\tau,h}^k(w_h) = \frac{1}{2\tau}\|w_h - u_h^{k-1}\|^2 + I(w_h)$$

in the set of functions $w_h \in \mathscr{S}^1(\mathscr{T}_h)$.

The scheme may be regarded as an implicit Euler method and is unconditionally stable.

Proposition 10.10 (Stability) *Assume that* $I : L^2(\Omega) \to \mathbb{R} \cup \{+\infty\}$ *is convex and lower-semicontinuous. For* $L = 1, 2, \dots, K$ *we have*

```
function tv_reg_regularized(d,red)
[c4n,n4e,Db,Nb] = triang_cube(d); c4n = c4n-.5;
for j = 1:red
    [c4n,n4e,Db,Nb] = red_refine(c4n,n4e,Db,Nb);
end
h = 2^(-red); alpha = 100; tau = h/10;
noise = .4; delta = h^(1/2);
nC = size(c4n,1); nE = size(n4e,1);
[¬,m,¬,¬] = fe_matrices(c4n,n4e);
gg = g(c4n)+noise*(rand(nC,1)-.5);
u = zeros(nC,1);
corr = 1; eps_stop = 1e-5;
while corr > eps_stop
    du = comp_gradient(c4n,n4e,u);
    a_du_inv = 1./sqrt(sum(du.^2,2)+delta^2);
    [s_du,¬] = fe_matrices_weighted(c4n,n4e,a_du_inv,zeros(nE,1));
    X = (1+alpha*tau)*m+tau*s_du;
    b = m*u+tau*alpha*m*gg;
    u_new = X\b;
    dt_u = (u_new-u)/tau;
    corr = sqrt(dt_u'*m*dt_u);
    u = u_new;
    show_p1(c4n,n4e,Db,Nb,u);
end

function val = g(x)
val = zeros(size(x,1),1);
val(sqrt(sum(x.^2,2))<.2) = 1;

function [s_a,m_b] = fe_matrices_weighted(c4n,n4e,a,b)
[nC,d] = size(c4n); nE = size(n4e,1);
m_loc = (ones(d+1,d+1)+eye(d+1))/((d+1)*(d+2));
ctr = 0; ctr_max = (d+1)^2*nE;
I = zeros(ctr_max,1); J = zeros(ctr_max,1);
X_s_a = zeros(ctr_max,1); X_m_b = zeros(ctr_max,1);
for j = 1:nE
    X_T = [ones(1,d+1);c4n(n4e(j,:),:)'];
    grads_T = X_T\[zeros(1,d);eye(d)];
    vol_T = det(X_T)/factorial(d);
    for m = 1:d+1
        for n = 1:d+1
            ctr = ctr+1;
            I(ctr) = n4e(j,m); J(ctr) = n4e(j,n);
            X_s_a(ctr) = vol_T*a(j)*grads_T(m,:)*grads_T(n,:)';
            X_m_b(ctr) = vol_T*b(j)*m_loc(m,n);
        end
    end
end
s_a = sparse(I,J,X_s_a,nC,nC); m_b = sparse(I,J,X_m_b,nC,nC);
```

Fig. 10.6 MATLAB realization of the semi-implicit gradient flow discretization of the regularized total variation functional I_δ defined in Algorithm 10.2

$$I(u_h^L) + \tau \sum_{k=1}^{L} \|d_t u_h^k\|^2 \le I(u_h^0).$$

Proof The existence of the iterates follows from the direct method in the calculus of variations, and the strong convexity of $I_{\tau,h}^k$ implies their uniqueness. For $k = 1, 2, \dots, K$ we have $0 \in \partial I_{\tau,h}^k(u_h^k)$, i.e., $-d_t u_h^k \in \partial I(u_h^k)$ and hence for all $v_h \in \mathscr{S}^1(\mathscr{T}_h)$

$$(-d_t u_h^k, v_h - u_h^k) + I(u_h^k) \le I(v_h).$$

The choice $v_h = u_h^{k-1}$ yields

$$\tau \|d_t u_h^k\|^2 + \tau d_t I(u_h^k) \le 0$$

and a summation over $k = 1, 2, \dots, L$ implies the stability estimate. □

We next bound the difference between the fully discrete and semi-discrete approximations, i.e., we estimate the difference $u_h^k - u^k$, where $(u^k)_{k=0,1,\dots,K}$ is the sequence of minimizers for the functionals

$$I_\tau^k(w) = \frac{1}{2\tau}\|w - u^{k-1}\|^2 + I(w)$$

with an initial $u^0 = u_0 \in L^2(\Omega)$. For ease of presentation we restrict to the case $I(u) = |Du|(\Omega)$.

Proposition 10.11 (Partial error estimate) *Let $I(u) = |Du|(\Omega)$ for $u \in BV(\Omega)$ and assume that $u_0 \in BV(\Omega) \cap L^\infty(\Omega)$. For $L = 1, 2, \dots, K$ we have*

$$\|u_h^L - u^L\|^2 \le \|u_h^0 - u_0\|^2 + ch^{1/3}.$$

The constant $c \ge 0$ depends on T, $|Du^0|(\Omega)$, $\|\nabla u_h^0\|_{L^1(\Omega)}$, and $\|u^0\|_{L^\infty(\Omega)}$.

Proof We let $(u^k)_{k=0,\dots,K} \subset BV(\Omega) \cap L^2(\Omega)$ be the solution of the semi-discrete scheme with initial value $u^0 = u_0$. Then, for $k = 1, 2, \dots, K$ and all $v \in BV(\Omega) \cap L^2(\Omega)$ we have

$$(-d_t u^k, v - u^k) + I(u^k) \le I(v).$$

For $k = 1, 2, \dots, K$, and all $v_h \in \mathscr{S}^1(\mathscr{T}_h)$ we have

$$(-d_t u_h^k, v_h - u_h^k) + I(u_h^k) \le I(v_h).$$

Choosing $v = u_h^k$ we deduce that

$$(d_t[u^k - u_h^k], u^k - u_h^k) + I(u^k) - I(v_h) \le (d_t u_h^k, v_h - u^k),$$

i.e.,

$$\frac{d_t}{2}\|u^k - u_h^k\|^2 + \frac{\tau}{2}\|d_t(u^k - u_h^k)\|^2 \le I(v_h) - I(u^k) + \|d_t u_h^k\|\|v_h - u^k\|.$$

For $\varepsilon > 0$ we let $v_h = u_{\varepsilon,h}^k$ be as in Lemma 10.1 so that

$$I(v_h) - I(u^k) \le c(\varepsilon + h\varepsilon^{-1})I(u^k)$$

and

$$\|v_h - u^k\|^2 \le \|v_h - u^k\|_{L^1(\Omega)}\|v_h - u^k\|_{L^\infty(\Omega)} \le c(h^2\varepsilon^{-1} + \varepsilon)|Du^k|(\Omega)\|u^k\|_{L^\infty(\Omega)}.$$

Arguing as in Proposition 10.2, we have $\|u^k\|_{L^\infty(\Omega)} \le \|u^0\|_{L^\infty(\Omega)}$ for $k = 1, 2, \dots, K$. The construction of $u_{\varepsilon,h}^k$ in Lemma 10.1 guarantees that $\|v_h\|_{L^\infty(\Omega)} \le \|u^k\|_{L^\infty(\Omega)}$. As in the proof of Proposition 10.10, we find that the semi-discrete scheme is energy-decreasing, i.e., we have $|Du^k|(\Omega) \le |Du^0|(\Omega)$ for $k = 1, 2, \dots, K$, and hence

$$|Du^k|(\Omega) + \tau\sum_{k=1}^{L}\|d_t u^k\|^2 \le |Du^0|(\Omega) = c_0.$$

Incorporating also the estimate from Proposition 10.10, it follows from a summation over $k = 1, 2, \dots, L$ that

$$\begin{aligned}\frac{1}{2}\|u_h^L - u^L\|^2 &\le \frac{1}{2}\|u_h^0 - u^0\|^2 + \tau\sum_{k=1}^{L}\big(|Dv_h|(\Omega) - |Du^k|(\Omega)\big)\\ &\quad + \Big(\tau\sum_{k=1}^{L}\|d_t u_h^k\|^2\Big)^{1/2}\Big(\tau\sum_{k=1}^{L}\|v_h - u^k\|^2\Big)^{1/2}\\ &\le \frac{1}{2}\|u_h^0 - u^0\|^2 + cT(\varepsilon + h\varepsilon^{-1})c_0\\ &\quad + cT^{1/2}c_0^{1/2}\|u^0\|_{L^\infty(\Omega)}^{1/2}(h^2\varepsilon^{-1} + \varepsilon)^{1/2}.\end{aligned}$$

Choosing $\varepsilon = h^{2/3}$ leads to the assertion. □

The combination of Proposition 10.11 with the abstract error estimate for implicit discretizations of subdifferential flows of Theorem 4.7 leads to the following error estimate.

Theorem 10.9 (Error estimate) *Assume that $u_0 \in BV(\Omega) \cap L^\infty(\Omega)$ and $u_h^0 \in \mathscr{S}^1(\mathscr{T}_h)$ is such that $\|u_0 - u_h^0\| \le h^{1/6}$ and $|Du_h^0|(\Omega) \le c$ for all $h > 0$. We then have*

$$\max_{k=1,\dots,K} \|u(t_k) - u_h^k\| \leq c(\tau^{1/2} + h^{1/6}).$$

Proof The assertion is a direct consequence of the abstract error estimate for implicit discretizations of subdifferential flows of Theorem 4.7 and Proposition 10.11. □

Remarks 10.17 (i) The upper bound can be improved to $\tau + h^{1/4}$ provided that $\partial I(u^0) \neq \emptyset$ and $\|d_t u_h^k\|_{L^\infty(\Omega)} \leq c$ for $k = 1, 2, \dots, K$.
(ii) In the case of Dirichlet boundary conditions and $d = 1$, any monotone function $u \in BV(\Omega)$ is stationary for I, whereas only the affine interpolant of the boundary data is stationary for the regularized functional I_δ.

10.3 Segmentation

We discuss in this section the numerical approximation of segmentation problems. The considered simple model problems detect edges in certain images and serve as bases for the development of models that describe damage and fracture in solid mechanics. We refer the reader to [5, 9] for further details.

10.3.1 The Mumford–Shah Functional

The *Mumford–Shah* functional detects certain edges in an image $g : \Omega \to \mathbb{R}$ by minimizing the functional

$$I(u, K) = \frac{\alpha}{2} \int_{\Omega \setminus K} |\nabla u|^2 \, dx + \beta \mathscr{H}^{d-1}(K) + \frac{\gamma}{2} \int_{\Omega \setminus K} (u - g)^2 \, dx$$

in closed sets $K \subset \overline{\Omega}$ and functions $u \in H^1(\Omega \setminus K)$ with given parameters $\alpha, \beta, \gamma > 0$. For a minimizing pair (u, K) the $(d-1)$-dimensional Hausdorff measure $\mathscr{H}^{d-1}(K)$ has to be finite, e.g., K is the union of curves or surfaces for $d = 2$ or $d = 3$, respectively, and $\mathscr{H}^{d-1}$ is the corresponding surface measure. The function u approximates the data g and may jump across the set K. Establishing the existence of minimizing pairs is a difficult task, since the unknowns u and K are different objects and the Hausdorff measure is not lower semicontinuous.

Example 10.6 For $k \in \mathbb{N}$ recursively define $S_k \subset [0, 1]$ through $S_0 = [0, 1/2]$ and

$$S_k = (1/2)S_{k-1} \cup (1/2)\big(S_{k-1} + 1/2\big) = \cup_{\ell=0}^{2^k-1} 2^{-(k+1)}[2\ell, 2\ell + 1]$$

e.g., $S_1 = [0, 1/4] \cup [2/4, 3/4]$. Then the sequence $(S_k)_{k \in \mathbb{N}}$ converges to $S = [0, 1]$ with respect to the Hausdorff metric

$$d_{\mathscr{H}}(K, L) = \inf\{\varepsilon > 0 : K \subset U_\varepsilon(L),\ L \subset U_\varepsilon(K)\},$$

where $U_\varepsilon(K) = \{x \in \mathbb{R}^d : \operatorname{dist}(x, K) < \varepsilon\}$. Since $\mathscr{H}^{d-1}(S) = 1$ and $\mathscr{H}^{d-1}(S_k) = 1/2$ for all $k \in \mathbb{N}$, we conclude that the mapping $K \mapsto \mathscr{H}^{d-1}(K)$ is not lower semicontinuous with respect to the Hausdorff metric.

The main idea to establish the existence of solutions is to consider functions of bounded variation and to identify K with the discontinuity set S_u of a function $u \in BV(\Omega)$. We recall that the distributional derivative of $u \in BV(\Omega)$ permits the decomposition

$$Du = \nabla u \otimes \mathrm{d}x - [\![un]\!] \otimes \mathrm{d}s|_{S_u} + C_u$$

with a vector field $\nabla u \in L^1(\Omega; \mathbb{R}^d)$ and the discontinuity set S_u of finite $(d-1)$-dimensional Hausdorff measure. The Cantor part C_u is in general supported on a set of infinite $(d-1)$-dimensional Hausdorff measure. If $C_u = 0$, it is natural to consider

$$I'(u) = \frac{\alpha}{2} \int_\Omega |\nabla u|^2 \,\mathrm{d}x + \beta \mathscr{H}^{d-1}(S_u) + \frac{\gamma}{2} \int_\Omega (u - g)^2 \,\mathrm{d}x.$$

The functions $u \in BV(\Omega)$ with $C_u = 0$ are called *special functions of bounded variation* and the set of all such functions is denoted $SBV(\Omega)$, i.e.,

$$SBV(\Omega) = \{u \in BV(\Omega) : C_u = 0\}.$$

Sequences $(u_j)_{j \in \mathbb{N}} \subset SBV(\Omega) \cap L^\infty(\Omega)$ that are uniformly bounded in $L^\infty(\Omega)$ and for which we have $\nabla u_j \in L^2(\Omega)$ for every $j \in \mathbb{N}$, such that the expression

$$\int_\Omega |\nabla u_j|^2 \,\mathrm{d}x + \mathscr{H}^{d-1}(S_{u_j})$$

is uniformly bounded, provide convergent subsequences $(u_{j_k})_{k \in \mathbb{N}}$ with limit $u \in SBV(\Omega)$, i.e., we have that $u_{j_k} \to u$ almost everywhere in Ω, $\nabla u_{j_k} \rightharpoonup \nabla u$ in $L^2(\Omega)$, and

$$\mathscr{H}^{d-1}(S_u) \le \liminf_{k \to \infty} \mathscr{H}^{d-1}(S_{u_{j_k}}).$$

This compactness property implies the following existence result.

Theorem 10.10 (Existence [1]) *If $g \in L^\infty(\Omega)$, then the functional I' has a minimizer $u \in SBV(\Omega) \cap L^\infty(\Omega)$. The pair (u, K) with $K = \overline{S_u} \cap \overline{\Omega}$ minimizes the Mumford–Shah functional in pairs (u, K) consisting of a closed set $K \subset \overline{\Omega}$ with $\mathscr{H}^{d-1}(K) < \infty$ and $u \in W^{1,2}(\Omega \backslash K)$.*

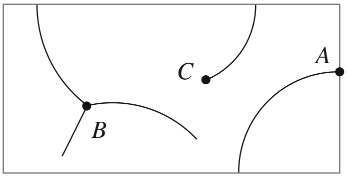


Fig. 10.7 Typical vertices of the singularity set K in the minimization of the Mumford–Shah functional; vertices are either points on the boundary where K intersects $\partial\Omega$ perpendicularly (A), triple points where three smooth segments intersect with equal angles (B), or endpoints of curves (C)

Precise characterizations of the singularity set K are available.

Remark 10.18 Assume $d = 2$ and a minimizing pair (u, K) is such that K is the finite union of $C^{1,1}$ curves. Then every vertex of K is either (a) A point on $\partial\Omega$ where K and $\partial\Omega$ intersect perpendicularly, (b) A point in Ω at which three $C^{1,1}$ curves intersect with angles $2\pi/3$, or (c) A point in Ω at which a $C^{1,1}$ curve ends, cf. Fig. 10.7. The technical results follow from contradictions and local modifications to lower the energy.

10.3.2 Regularization of $I'(u)$

It is difficult to approximate the Mumford–Shah functional directly with finite element methods since the singularity sets of discontinuous, piecewise polynomial finite element functions are subsets of the skeleton of the underlying triangulation which is in general too restrictive to approximate a given curve. An approach to regularizing the Mumford–Shah functional is to describe the set K by the zero level set $\Gamma_\phi = \phi^{-1}(\{0\})$ of a function $\phi : \Omega \to \mathbb{R}$ and noting that the Hausdorff measure of Γ_ϕ is approximated by the Modica–Mortola type length functional L_ε, i.e.,

$$\mathscr{H}^{d-1}(\Gamma_\phi) \approx L_\varepsilon(\Gamma_\phi) = \frac{\varepsilon}{2}\int_\Omega |\nabla\phi|^2 \,\mathrm{d}x + \frac{1}{2\varepsilon}\int_\Omega (\phi - 1)^2 \,\mathrm{d}x.$$

This relation follows from Young's inequality together with the transformation $w = (\phi - 1)^2$, i.e., $|\nabla w| = 2|\phi - 1||\nabla\phi|$. We have

$$L_\varepsilon(\Gamma_\phi) = \frac{\varepsilon}{2}\int_\Omega |\nabla\phi|^2 \,\mathrm{d}x + \frac{1}{2\varepsilon}\int_\Omega (\phi-1)^2 \,\mathrm{d}x \geq \int_\Omega |\nabla\phi||\phi-1| \,\mathrm{d}x = \frac{1}{2}\int_\Omega |\nabla w| \,\mathrm{d}x.$$

We assume that Γ_ϕ is a smooth curve and, for every $r \in \Gamma_\phi$, denote by n_r the unit normal to Γ_ϕ at r. With the tubular neighborhood

$$\Gamma_{\phi,\varepsilon} = \{x \in \Omega : x = r + tn_r, |t| \le \varepsilon\}$$

of Γ_ϕ we have

$$L_\varepsilon(\Gamma_\phi) \ge \frac{1}{2}\int_{\Gamma_{\phi,\varepsilon}} |\nabla w| \,\mathrm{d}x \ge \frac{1}{2}\int_{\Gamma_\phi}\int_{-\varepsilon}^{\varepsilon} |\nabla w \cdot n_r| \,\mathrm{d}t \,\mathrm{d}r.$$

Assuming that $L_\varepsilon(\Gamma_\phi)$ remains bounded as $\varepsilon \to 0$, the function ϕ approaches the value 1 away from Γ_ϕ for ε sufficiently small, so that we may assume that $w = (\phi - 1)^2 \approx 0$ in $\Omega \setminus \Gamma_{\phi,\varepsilon}$. The integral of the modulus of the derivative of w in normal direction to Γ_ϕ is then approximately 2 and we obtain

$$L_\varepsilon(\Gamma_\phi) \ge \int_{\Gamma_\phi} 1 \,\mathrm{d}s = \mathscr{H}^{d-1}(\Gamma_\phi).$$

These observations motivate us to consider the Ambrosio–Tortorelli approximation of the Mumford–Shah functional in which L_ε approximates $\mathscr{H}^{d-1}(S_u)$ and enforces ϕ to be close to one, while a term $\phi^2|\nabla u|^2$ favors $\phi \approx 0$ to permit large, unbounded gradients of u.

Theorem 10.11 (Regularization [3]) *For* $(u, \phi) \in H^1(\Omega) \times H^1(\Omega)$ *and* $\varepsilon > 0$, *define the* Ambrosio–Tortorelli functional

$$\begin{aligned} AT_\varepsilon(u,\phi) = {} & \frac{\alpha}{2}\int_\Omega (\phi^2 + \varepsilon^2)|\nabla u|^2 \,\mathrm{d}x \\ & + \beta\Big(\frac{\varepsilon}{2}\int_\Omega |\nabla\phi|^2 \,\mathrm{d}x + \frac{1}{2\varepsilon}\int_\Omega (\phi - 1)^2 \,\mathrm{d}x\Big) + \frac{\gamma}{2}\int_\Omega (u - g)^2 \,\mathrm{d}x \end{aligned}$$

and extend AT_ε *with value* $+\infty$ *to* $L^1(\Omega) \times L^1(\Omega)$. *Then, as* $\varepsilon \to 0$, *we have that* $AT_\varepsilon \to^\Gamma I''$ *with respect to strong convergence in* $L^1(\Omega) \times L^1(\Omega)$, *and where* $I''(u,\phi) = I'(u)$ *if* $(u,\phi) \in SBV(\Omega) \times L^1(\Omega)$ *with* $\phi = 1$ *almost everywhere and* $I''(u,\phi) = +\infty$ *otherwise, i.e.,* $I'(u) = I''(u, 1)$ *for all* $u \in SBV(\Omega)$.

10.3.3 Numerical Approximation of AT_ε

The functional AT_ε can be directly discretized with H^1-conforming finite element methods; that is, given $\varepsilon > 0$ and a triangulation $\mathscr{T}_h$ of Ω, we consider the separately convex functional

$$AT_{\varepsilon,h}(u_h,\phi_h) = \frac{\alpha}{2}\int_\Omega (\phi_h^2+\varepsilon^2)|\nabla u_h|^2\,\mathrm{d}x \\ + \beta\Big(\frac{\varepsilon}{2}\int_\Omega |\nabla\phi_h|^2\,\mathrm{d}x + \frac{1}{2\varepsilon}\int_\Omega (\phi_h-1)^2\,\mathrm{d}x\Big) + \frac{\gamma}{2}\int_\Omega (u_h-g)^2\,\mathrm{d}x$$

for $(u_h,\phi_h)\in\mathscr{S}^1(\mathscr{T}_h)$. Extending $AT_{\varepsilon,h}$ by $+\infty$ on $L^1(\Omega)^2\setminus\mathscr{S}^1(\mathscr{T}_h)^2$, the density of $\mathscr{S}^1(\mathscr{T}_h)$ in $L^1(\Omega)$ leads to a Γ-convergence result as in Theorem 10.11. The iterative solution of $AT_{\varepsilon,h}$ is based on a semi-implicit discretization of a gradient flow with respect to ϕ_h. This leads to two uncoupled equations in every step of the iteration. We let $P_0 v\in\mathscr{L}^0(\mathscr{T}_h)$ denote the elementwise average of a function $v\in L^1(\Omega)$.

Algorithm 10.4 (*Semi-implicit gradient flow for $AT_{\varepsilon,h}$*) Given $\tau>0$ and $\phi_h^0\in\mathscr{S}^1(\mathscr{T}_h)$, define the sequence $(u_h^k,\phi_h^k)_{k=1,2,\ldots}$ by solving for $k=1,2,\ldots$ the equations

$$\alpha((|P_0\phi_h^{k-1}|^2+\varepsilon^2)\nabla u_h^k,\nabla v_h)+\gamma(u_h^k-g,v_h)=0,$$
$$(d_t\phi_h^k,w_h)+\alpha(|\nabla u_h^k|^2\phi_h^k,w_h)+\beta\varepsilon(\nabla\phi_h^k,\nabla w_h)+\frac{\beta}{\varepsilon}(\phi_h^k-1,w_h)=0$$

for all $(v_h,w_h)\in\mathscr{S}^1(\mathscr{T}_h)\times\mathscr{S}^1(\mathscr{T}_h)$. Stop the iteration if $\|d_t\phi_h^k\|\le\varepsilon_{\mathrm{stop}}$.

In the implementation of the scheme shown in Fig. 10.8 we used the parameter $\beta=1$.

10.3.4 The Perona–Malik Equation

The *Perona–Malik* equation is a nonlinear parabolic partial differential equation that denoises an image g for a parameter $\lambda>0$ through

$$\partial_t u - \operatorname{div}\Big(\frac{\nabla u}{(1+|\nabla u|^2/\lambda^2)^2}\Big)=0,\quad \partial_n u(t,\cdot)=0,\quad u(0)=g.$$

The diffusion coefficient $a(|\nabla u|)=(1+|\nabla u|^2/\lambda^2)^{-2}$ is small in regions where $|\nabla u|$ is large and this leads to a preservation of edges in the images that are characterized by large gradients. In the remaining regions where $|\nabla u|\le c$, the diffusion coefficient $a(|\nabla u|)$ is larger and causes a smoothing of u away from the edges. This leads to a simultaneous denoising and steepening of edges, but analytically to the problem that the equation is of backward and forward parabolic type, so that the well-posedness of the initial boundary value problem is false in general. The equation has an interesting relation to the Mumford–Shah model, i.e., to its Ambrosio–Tortorelli regularization, described in [13]. An implicit discretization in time of the Perona–Malik equation leads to the problem of determining u^k such that

```
function ambrosio_tortorelli(d,red)
[c4n,n4e,Db,Nb] = triang_cube(d); c4n = 2*(c4n-.5);
for j = 1:red
    [c4n,n4e,Db,Nb] = red_refine(c4n,n4e,Db,Nb);
end
[nC,d] = size(c4n); nE = size(n4e,1); gg = g(c4n);
alpha = 1; gamma = 10; tau = 2^(-red)/10; eps = 1/10;
[s,m,¬] = fe_matrices(c4n,n4e);
a_0 = zeros(nE,1);
phi = zeros(nC,1); corr = 1; eps_stop = 1e-2;
while corr > eps_stop
    a_phi_sq = eps^2+(sum(phi(n4e),2)/(d+1)).^2;
    [s_phi,¬] = fe_matrices_weighted(c4n,n4e,a_phi_sq,a_0);
    X_u = gamma*m+alpha*s_phi;
    b_u = gamma*m*gg;
    u = X_u\b_u;
    du = comp_gradient(c4n,n4e,u);
    mod_du_sq = sum(du.^2,2);
    [¬,m_du] = fe_matrices_weighted(c4n,n4e,a_0,mod_du_sq);
    X_phi = m+eps*tau*s+tau*alpha*m_du+(1/(2*eps))*tau*m;
    b_phi = m*phi+(1/(2*eps))*tau*m*ones(nC,1);
    phi_new = X_phi\b_phi;
    dt_phi = (phi_new-phi)/tau;
    corr = sqrt(dt_phi'*m*dt_phi);
    phi = phi_new;
    figure(1); show_p1(c4n,n4e,Db,Nb,u);
    figure(2); show_p1(c4n,n4e,Db,Nb,phi);
end

function val = g(x)
val = tanh(100*(sum(x.^2,2)-1/2));
```

Fig. 10.8 MATLAB realization of Algorithm 10.4 for the iterative minimization of the Ambrosio–Tortorelli regularization of the Mumford–Shah functional

$$\operatorname{div}\left(\frac{\nabla u^k}{(1+|\nabla u^k|^2/\lambda^2)^2}\right) = \frac{1}{\tau}(u^k - u^{k-1}). \tag{10.1}$$

The Euler–Lagrange equations of the Ambrosio–Tortorelli functional AT_ε define the pair (u,ϕ) via

$$\alpha \operatorname{div}\big((\phi^2+\varepsilon^2)\nabla u\big) = \gamma(u-g),$$
$$\alpha\varepsilon|\nabla u|^2\phi - \beta\varepsilon^2\Delta\phi + \beta(\phi-1) = 0.$$

Neglecting terms with a factor ε^2, we find that

$$\phi = \frac{1}{1+(\alpha/\beta)\varepsilon|\nabla u|^2}$$

and

$$\operatorname{div}\left(\frac{\nabla u}{\left(1+(\alpha/\beta)\varepsilon|\nabla u|^2\right)^2}\right) = \frac{\gamma}{\alpha}(u-g). \tag{10.2}$$

For $k = 1$ and $u^0 = g$ in (10.1) and, e.g., $\alpha = \lambda^{-1/2}$, $\beta = \varepsilon$, and $\gamma = \alpha/\tau$ in (10.2), the partial differential equations coincide. The practical solution of the Perona–Malik equation is based on a semi-implicit discretization of the equation.

Algorithm 10.5 (*Semi-implicit Perona–Malik equation*) Given $\tau > 0$ and $g_h \in \mathscr{S}^1(\mathscr{T}_h)$, define the sequence $(u_h^k)_{k=0,1,\dots}$ by setting $u_h^0 = g_h$ and solving for $k = 1, 2, \dots$ the equations

$$(d_t u_h^k, v_h) + \left(\frac{\nabla u_h^k}{(1+|\nabla u_h^{k-1}|^2/\lambda^2)^2}, \nabla v_h\right) = 0$$

for all $v_h \in \mathscr{S}^1(\mathscr{T}_h)$. Stop the iteration if $\|d_t u_h^k\| \le \varepsilon_{\text{stop}}$.

An implementation of the scheme is shown in Fig. 10.9.

```
function perona_malik(d,red)
[c4n,n4e,Db,Nb] = triang_cube(d); c4n = 2*(c4n-.5);
lambda = .5;
for j = 1:red
    [c4n,n4e,Db,Nb] = red_refine(c4n,n4e,Db,Nb);
end
nE = size(n4e,1);
tau = 2^(-red)/10;
[¬,m,¬] = fe_matrices(c4n,n4e);
u = g(c4n);
corr = 1; eps_stop = 1e-2;
while corr > eps_stop
    du = comp_gradient(c4n,n4e,u);
    a_du = (1+sum(du.^2,2)/lambda^2).^(-2);
    [s_du,¬] = fe_matrices_weighted(c4n,n4e,a_du,zeros(nE,1));
    X = m+tau*s_du;
    b = m*u;
    u_new = X\b;
    dt_u = (u_new-u)/tau;
    u = u_new;
    corr = sqrt(dt_u'*m*dt_u);
    show_p1(c4n,n4e,Db,Nb,u);
end

function val = g(x)
val = tanh(100*(sum(x.^2,2)-1/2))+.25*(rand(size(x,1),1)-.5);
```

Fig. 10.9 MATLAB realization of the semi-implicit discretization of the Perona–Malik equation specified in Algorithm 10.5

Remarks 10.19 (i) A stability proof for the iteration is expected to require restrictive conditions on the step size τ. Practically, the iteration provides satisfactory results for $\tau \leq ch$. Difficulties in the numerical analysis reflect the fact that no general existence theory for the Perona–Malik equation is available and in fact solutions may fail to exist due to occurring backward diffusion.

(ii) An alternative choice for the diffusion coefficient in the Perona–Malik equation is $a(s) = e^{-s^2/\lambda^2}$.

References

1. Ambrosio, L.: Existence theory for a new class of variational problems. Arch. Ration. Mech. Anal. **111**(4), 291–322 (1990). http://dx.doi.org/10.1007/BF00376024
2. Ambrosio, L., Fusco, N., Pallara, D.: Functions of Bounded Variation and Free Discontinuity Problems. Oxford Mathematical Monographs, The Clarendon Press. Oxford University Press, New York (2000)
3. Ambrosio, L., Tortorelli, V.M.: Approximation of functionals depending on jumps by elliptic functionals via Γ-convergence. Commun. Pure Appl. Math. **43**(8), 999–1036 (1990). http://dx.doi.org/10.1002/cpa.3160430805
4. Attouch, H., Buttazzo, G., Michaille, G.: Variational Analysis in Sobolev and BV Spaces. MPS/SIAM Series on Optimization, vol. 6. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2006)
5. Aubert, G.: Mathematical Problems in Image Processing. Applied Mathematical Sciences, vol. 147, 2nd edn. Springer, New York (2006)
6. Bartels, S.: Total variation minimization with finite elements: convergence and iterative solution. SIAM J. Numer. Anal. **50**(3), 1162–1180 (2012). http://dx.doi.org/10.1137/11083277X
7. Bartels, S.: Broken Sobolev space iteration for total variation regularized minimization problems (2013). Preprint
8. Bartels, S., Nochetto, R.H., Salgado, A.J.: Discrete total variation flows without regularization. SIAM J. Numer. Anal. **52**(1), 363–385 (2014). http://dx.doi.org/10.1137/120901544
9. Braides, A.: Approximation of Free-Discontinuity Problems. Lecture Notes in Mathematics, vol. 1694. Springer, Berlin (1998)
10. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011). http://dx.doi.org/10.1007/s10851-010-0251-1
11. Ekeland, I., Témam, R.: Convex Analysis and Variational Problems. Classics in Applied Mathematics, vol. 28, English edn. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1999). http://dx.doi.org/10.1137/1.9781611971088
12. Hintermüller, M., Kunisch, K.: Total bounded variation regularization as a bilaterally constrained optimization problem. SIAM J. Appl. Math. **64**(4), 1311–1333 (2004). http://dx.doi.org/10.1137/S0036139903422784
13. Kawohl, B.: From Mumford-Shah to Perona-Malik in image processing. Math. Methods Appl. Sci. **27**(15), 1803–1814 (2004). http://dx.doi.org/10.1002/mma.564
14. Rockafellar, R.T.: Convex Analysis. Princeton Mathematical Series, vol. 28. Princeton University Press, Princeton (1970)

Chapter 11
Elastoplasticity

11.1 Modeling and Analytical Properties

We discuss in this section the mathematical description of elastoplastic material behavior. We follow the textbooks [5, 9] and the survey article [7].

11.1.1 One-Dimensional Plastic Effects

Mathematical elasticity is based on the Hookean principle that a deformation of an elastic body is accompanied by a restoring interior force that pulls the body back into its reference configuration when an outer force stops acting. The starting point of elastoplasticity is that only a bounded range of restoring forces are possible, and when a limit is reached, microstructural changes in the crystal lattice occur that lead to remaining, plastic deformations. A typical example is the elongation of a rubber band or copper wire beyond a critical value that leads to a permanent lengthening of the band or wire. To specify some basic principles, we consider a one-dimensional wire that is regarded as a chain of elements consisting of springs and frictional devices, as depicted in Fig. 11.1.

If one end of the band is fixed and the other end is displaced, a change of length occurs, which causes a restoring force that is proportional to the relative change in length, i.e., to the strain, as long as it is below the friction coefficient. When the restoring force reaches the value of the friction coefficient, the frictional device starts to glide and a plastic strain compensates the increasing total strain while the stress remains constant. The rate of change of the plastic strain has the same sign as the stress.

Example 11.1 Suppose that a wire occupies the region $\Omega = (0, 1)$ and its left end is fixed while its right end is displaced gradually by $u_{\mathrm{D}}(t)$. The deformation of the wire is then given by $y(t, x) = x(1 + u_{\mathrm{D}}(t))$. The corresponding displacement is $u(t, x) = x u_{\mathrm{D}}(t)$ and the strain is given by $\partial_x u(t, x) = u_{\mathrm{D}}(t)$. The restoring force or stress σ is proportional to $\partial_x u$, i.e., $\sigma = \mathbb{C}\partial_x u$ as long as the magnitude

S. Bartels, *Numerical Methods for Nonlinear Partial Differential Equations*,
Springer Series in Computational Mathematics 47,
DOI 10.1007/978-3-319-13797-1_11

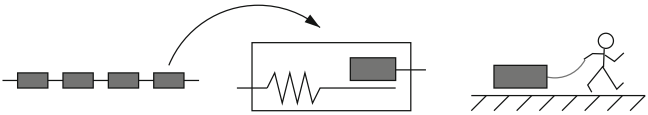


Fig. 11.1 Plastic material behavior interpreted as a combination of springs and frictional devices (*left*); an analogy is the pulling of a mass over a dry surface with an elastic rope (*right*)

of σ is below the critical value σ_y, i.e., $|\sigma| \le \sigma_y$. When this value is reached, a plastic strain develops while the stress remains constant, i.e., $\sigma = \mathbb{C}(\partial_x u - p)$ and $p = \max\{0, u_{\mathrm{D}}(t) - \mathbb{C}^{-1}\sigma_y\}$. The evolution of p is described by the requirements that $\dot{p} = 0$ as long as $|\sigma| < \sigma_y$ and $\dot{p}$ is proportional to σ when $|\sigma| = \sigma_y$.

11.1.2 Hypotheses of Multi-dimensional Elastoplasticity

We consider an object made of a metal or more generally of a ductile material that occupies the domain $\Omega \subset \mathbb{R}^3$ on which a body force $f : [0, T] \times \Omega :\to \mathbb{R}^3$ and a surface traction $g : [0, T] \times \Gamma_{\mathrm{N}} \to \mathbb{R}^3$ are acting. The corresponding displacement $u : \Omega \to \mathbb{R}^3$ is required to vanish on the boundary $\Gamma_{\mathrm{D}} = \partial\Omega \setminus \Gamma_{\mathrm{N}}$. Assuming that only small deformations occur, these can be described by the symmetric gradient $\varepsilon(u) = (\nabla u^\top + \nabla u)/2$ called *strain*. The corresponding restoring force is denoted by the symmetric stress tensor $\sigma \in \mathbb{R}^{3\times 3}_{\mathrm{sym}}$, and as long as σ belongs to a set of admissible forces, we have the linear relation $\sigma = \mathbb{C}\varepsilon(u)$. In a quasi-stationary situation, we have the equilibrium of forces

$$-\operatorname{div}\sigma = f \text{ in } \Omega, \quad \sigma n = g \text{ on } \Gamma_{\mathrm{N}}.$$

When strains occur that lead to inadmissible stresses, another variable is required and this is the *plastic strain*

$$p = \varepsilon(u) - \mathbb{C}^{-1}\sigma = \varepsilon(u) - e.$$

The variable p is a symmetric tensor and assuming that uniform compressions are entirely elastic, one imposes the plastic incompressibility condition that p is trace-free. When plastic material behavior occurs, the material properties often change, and this is described by an internal variable $\xi \in \mathbb{R}^m$, e.g., the proportionality relation between stress and strain may change. In particular, it is observed that the set of admissible stresses increases when plasticity occurs or that the center of the set of *admissible stresses* is shifted. These effects are called *isotropic* and *kinematic hardening*, respectively. Situations in which no hardening occurs and the set of admissible stresses remains unchanged are referred to as *perfect plasticity*. Considering an isothermal and rate-independent situation, the fundamental laws of thermodynamics allow us to deduce the existence of a free energy of the form

$$\phi(e, \xi) = \phi^e(e) + \phi^p(\xi).$$

Moreover, the additive decomposition $\varepsilon(u) = e + p$ can be justified from thermodynamical considerations. In the simplest linear setting we may assume that

$$\phi(e, \xi) = \frac{1}{2} e : \mathbb{C} e + \frac{1}{2} \xi : \mathbb{H} \xi$$

with symmetric and positive definite tensors $\mathbb{C} : \mathbb{R}^{d \times d}_{\mathrm{sym}} \to \mathbb{R}^{d \times d}_{\mathrm{sym}}$ and $\mathbb{H} : \mathbb{R}^m \to \mathbb{R}^m$. With this, the *stress tensor* is defined as $\sigma = \partial_e \phi(e, \xi) = \mathbb{C} e$ and we define the conjugate forces $\chi = -\partial_\xi \phi(e, \xi) = -\mathbb{H} \xi$. The pairs $\Sigma = (\sigma, \chi)$ and $P = (p, \xi)$ are called *generalized stress* and *generalized plastic strain*, respectively. The hypothesis of maximal plastic work reads as

$$\Sigma \cdot \dot{P} \geq T \cdot \dot{P}$$

for all admissible generalized stresses $T \in S$. This is equivalent to the Prandtl–Reuss normality rule or *flow rule* $\dot{P} \in N_S(\Sigma)$ with the normal cone $N_S(\Sigma)$ of the set S at Σ. In particular, the rate of change of the plastic strain vanishes if Σ belongs to the interior of S called *elastic domain*. The boundary of S is called the *yield surface*. A *yield function* is a function Φ that defines the set of admissible stresses as $S = \{\Sigma : \Phi(\Sigma) \leq 0\}$ and determines the yield surface as the zero level set of Φ. The modeling of a yield function is typically based on the formulation of a yield criterion that determines when plastic material behavior sets in and popular choices are the von Mises and the Tresca criteria, which model that plasticity occurs when certain shear stresses exceed a given threshold parameter.

Figure 11.2 illustrates different plasticity models by corresponding hysteresis curves, i.e., stress-strain relations, in a cyclic loading-unloading experiment. Up to time t_1, the strain ε increases and the stress σ is proportional to ε until the *yield stress* σ_y is reached. Then plastic material behavior occurs and while the strain increases,

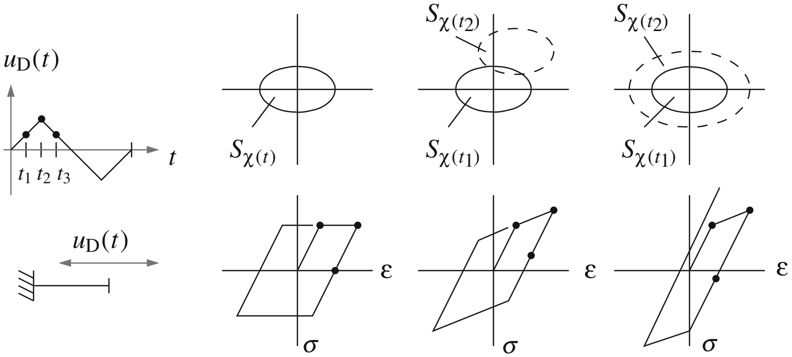


Fig. 11.2 Sets of admissible stresses $S_{\chi(t)} = \{\sigma \in \mathbb{R}^{d \times d}_{\mathrm{sym}} : \Phi(\sigma, \chi(t)) \leq 0\}$ for given internal forces $\chi(t)$ and hysteresis curves for different hardening models in a cyclic loading-unloading experiment; the stress-strain relations show different hysteresis effects for perfect plasticity (*second column*), kinematic hardening (*third column*), and isotropic hardening (*fourth column*)

the stress remains constant in the case of perfect plasticity or continuous to increase with a different rate in the case of kinematic or isotropic hardening. This is accompanied by a change of the internal variable or equivalently the set of admissible stresses. When the direction of loading changes, an elastic unloading takes place until the boundary of the modified set of admissible stresses is reached. Practically, the experiment is carried out by extending a thin wire by a prescribed amount and measuring the required force.

11.1.3 Mathematical Model

Based on the previous discussion we formulate the isothermal, quasi-static elastoplastic model problem with Prandtl–Reuss flow rule. For a bounded domain $\Omega \subset \mathbb{R}^d$, $\Gamma_{\mathrm{D}} \subset \partial\Omega$, $\Gamma_{\mathrm{N}} = \partial\Omega \setminus \Gamma_{\mathrm{D}}$ and $f : [0,T] \times \Omega \to \mathbb{R}^d$ and $g : [0,T] \times \Gamma_{\mathrm{N}} \to \mathbb{R}^d$, we seek $(u, p, \xi) : [0,T] \times \Omega \to \mathbb{R}^d \times \mathbb{R}^{d\times d}_{\mathrm{sym}} \times \mathbb{R}^m$ with $(u, p, \xi)(0) = (u_0, p_0, \xi_0)$ such that

$$\begin{aligned} -\operatorname{div}\sigma &= f, & (\dot{p}, \dot{\xi}) &\in \partial I_S(\sigma, \chi), \\ \sigma n|_{\Gamma_{\mathrm{N}}} &= g, & \sigma &= \mathbb{C}(\varepsilon(u) - p), \\ u|_{\Gamma_{\mathrm{D}}} &= 0, & \chi &= -\mathbb{H}\xi. \end{aligned}$$

Inhomogeneous Dirichlet boundary conditions are assumed to be included in the right-hand side. The subdifferential of the indicator functional I_S of S evaluated at Σ coincides with the normal cone $N_S(\Sigma)$ and the condition $\dot{P} \in \partial I_S(\Sigma)$ is equivalent to $\Sigma \in \partial I_S^*(\dot{P})$ with the support functional I_S^* of S. The inclusion can thus be equivalently formulated by requiring that

$$\sigma : (q - \dot{p}) + \chi \cdot (\zeta - \dot{\xi}) + I_S^*\big(\dot{p}, \dot{\xi}\big) \le I_S^*(q, \zeta)$$

is satisfied for all $(q, \zeta) \in \mathbb{R}^{d\times d}_{\mathrm{sym}} \times \mathbb{R}^m$. To derive a weak formulation, we set

$$Y = H^1_{\mathrm{D}}(\Omega; \mathbb{R}^d) \times L^2(\Omega; \mathbb{R}^{d\times d}_{\mathrm{sym}}) \times L^2(\Omega; \mathbb{R}^m)$$

and define the bilinear form $\mathscr{A} : Y \times Y \to \mathbb{R}$ for $y = (u, p, \xi)$ and $w = (v, q, \zeta)$ by

$$\mathscr{A}(y, w) = \int_\Omega \mathbb{C}(\varepsilon(u) - p) : (\varepsilon(v) - q) + \mathbb{H}\xi \cdot \zeta \,\mathrm{d}x$$

and the linear form

$$\ell(w) = \int_\Omega f \cdot v \,\mathrm{d}x + \int_{\Gamma_{\mathrm{N}}} g \cdot v \,\mathrm{d}s.$$

The variational inequality can be written as $(\sigma, \chi) \in \partial\psi(\dot{p}, \dot{\xi})$ with the *dissipation functional* $\psi : Y \to \mathbb{R} \cup \{+\infty\}$ defined for $w = (v, q, \zeta)$ by

$$\psi(w) = \int_\Omega I_S^*(q, \zeta)\, \mathrm{d}x.$$

The model problem is now formally equivalent to finding $y : [0, T] \to Y$ such that $y(0) = y_0$ and

$$\mathscr{A}(y, w - \dot{y}) + \psi(w) - \psi(\dot{z}) \geq \ell(w - \dot{y})$$

for all $w \in Y$ and all $t \in [0, T]$. A proof follows from choosing $w = (\pm v + \dot{u}, 0, 0)$ to deduce the weak formulation of the equilibrium of forces and $w = (\dot{u}, q, \xi)$ to verify a weak form of the flow rule.

11.1.4 Flow Rules and Coercivity

A yield function that realizes the *von Mises yield criterion* and describes kinematic and isotropic hardening simultaneously is given by

$$\Phi(\sigma, \alpha, \beta) = |\operatorname{dev}(\sigma + \beta)| - \sigma_y(1 + \alpha_+)$$

with $\operatorname{dev}\sigma = \sigma - (1/d)\operatorname{tr}\sigma$ and $\alpha_+ = \max\{\alpha, 0\}$ for a generalized stress vector $\Sigma = (\sigma, \alpha, \beta) \in \mathbb{R}^{d\times d}_{\mathrm{sym}} \times \mathbb{R} \times \mathbb{R}^{d\times d}_{\mathrm{sym}}$ with a yield stress $\sigma_y > 0$. The set of admissible stresses is defined as

$$S = \{\Sigma \in \mathbb{R}^{d\times d}_{\mathrm{sym}} \times \mathbb{R} \times \mathbb{R}^{d\times d}_{\mathrm{sym}} : \Phi(\Sigma) \leq 0\}.$$

Note that here the internal variable ξ is identified with the pair $(a, b) \in \mathbb{R} \times \mathbb{R}^{m\times m}_{\mathrm{sym}}$ and the variable χ is given by $(\alpha, \beta) = -\mathbb{H}(a, b)$. The support functional I_S^* for S can be computed explicitly.

Lemma 11.1 (General support functional) *For* $\dot{P} = (\dot{p}, \dot{a}, \dot{b}) \in \mathbb{R}^{d\times d}_{\mathrm{sym}} \times \mathbb{R} \times \mathbb{R}^{d\times d}_{\mathrm{sym}}$ *we have*

$$I_S^*(\dot{P}) = \begin{cases} \sigma_y|\dot{p}| & \text{if } \operatorname{tr}\dot{p} = 0,\ \dot{b} = \dot{p},\ \sigma_y|\dot{p}| \leq -\dot{a}, \\ +\infty & \text{otherwise.} \end{cases}$$

Proof By definition of $I_S^*(\dot{P})$ we have

$$I_S^*(\dot{p}, \dot{a}, \dot{b}) = \sup_{(\sigma,\alpha,\beta)\in S} \dot{p} : \sigma + \dot{a}\,\alpha + \dot{b} : \beta.$$

If $\operatorname{tr}\dot{p} \neq 0$, we choose $(\sigma, \alpha, \beta) = (rI_d, 0, 0)$, i.e., $\operatorname{dev}\sigma = 0$, for arbitrary $r \in \mathbb{R}$ and deduce that $I_S^*(\dot{p}, \dot{a}, \dot{b}) = \infty$. If $\dot{p} \neq \dot{b}$, we choose $\sigma = r(\dot{p} - \dot{b})$, $\beta = -\sigma$, and $\alpha = 0$ to deduce that $I_S^*(\dot{p}, \dot{a}, \dot{b}) = \infty$. If $\sigma_y|\dot{p}| > -\dot{a}$, we choose $\sigma = \sigma_y(1+r)\dot{p}/|\dot{p}|$, $\beta = 0$, and $\alpha = r$ for $r \geq 0$ so that

$$I_S^*(\dot{p}, \dot{a}, \dot{b}) \geq \sigma_y(1+r)|\dot{p}| + \dot{a}r \geq \sigma_y|\dot{p}| + (\sigma_y|\dot{p}| + \dot{a})r$$

which is unbounded as $r \to \infty$. We may thus assume $\dot{p} = \dot{b}$ and $\operatorname{tr} \dot{p} = 0$, i.e., $\dot{p} = \operatorname{dev} \dot{p}$, and $\sigma_y|\dot{p}| \leq -\dot{a}$ in the following. For every $\alpha \geq 0$, the maximal trace-free choice for σ and β is given by $\sigma = -\beta = (1+\alpha)(\sigma_y/2)\dot{p}/|\dot{p}|$ so that

$$I_S(\dot{p}, \dot{a}, \dot{b}) = \sup_{\alpha \geq 0}(1+\alpha)\sigma_y|\dot{p}| + \alpha\dot{a} = \sup_{\alpha \geq 0} \sigma_y|\dot{p}| + (\sigma_y|\dot{p}| + \dot{a})\alpha = \sigma_y|\dot{p}|$$

since $\sigma_y|\dot{p}| + \dot{a} \leq 0$. □

Special cases of the flow rule are the following.

Examples 11.2 (i) Perfect plasticity corresponds to

$$\Phi(\sigma) = |\operatorname{dev}(\sigma)| - \sigma_y$$

and the variables (α, β) and (a, b) can be eliminated from the problem.
(ii) Linear isotropic hardening corresponds to

$$\Phi(\sigma, \alpha) = |\operatorname{dev}(\sigma)| - \sigma_y(1 + \alpha_+)$$

and the variables β and b can be eliminated from the problem.
(iii) Kinematic hardening corresponds to

$$\Phi(\sigma, \beta) = |\operatorname{dev}(\sigma + \beta)| - \sigma_y$$

and the variables α and a can be eliminated from the problem. The variable $\beta = -\mathbb{H}_{\mathrm{kin}}b$ is called back stress and can be replaced by $-\mathbb{H}_{\mathrm{kin}}p$, noting that $\dot{p} = \dot{b}$ on dom ψ and assuming $p(0) = b(0)$.

In the case of linear kinematic or isotropic hardening, we have that the bilinear form $\mathscr{A}$ is coercive on the domain of ψ, i.e., on

$$\operatorname{dom} \psi = \{w \in Y : \psi(w) < \infty\},$$

where $\psi(w) = \int_\Omega I_S^*(q, \zeta)\,\mathrm{d}x$ for $w = (v, q, \zeta) \in Y$.

Proposition 11.1 (Coercivity) *Assume that for $\xi = (a, b) \in \mathbb{R} \times \mathbb{R}^{d \times d}_{\mathrm{sym}}$, we have*

$$\mathbb{H}\xi : \xi = \mathbb{H}_{\mathrm{iso}}a^2 + \mathbb{H}_{\mathrm{kin}}b : b$$

such that either $\mathbb{H}_{\mathrm{iso}}$ or $\mathbb{H}_{\mathrm{kin}}$ is positive definite. If $H_{\mathrm{iso}} = 0$ or $\mathbb{H}_{\mathrm{kin}} = 0$, then a or b is eliminated from the problem, respectively. If $\mathbb{H}_{\mathrm{kin}} \neq 0$, then the variables p and *b are identified. With this convention the bilinear form*

$$\mathscr{A}(y,w) = \frac{1}{2}\int\limits_{\Omega} \mathbb{C}\big(\varepsilon(u)-p\big) : \big(\varepsilon(v)-q\big)\,\mathrm{d}x + \frac{1}{2}\int\limits_{\Omega} a\mathbb{H}_{\text{iso}}e + b : \mathbb{H}_{\text{kin}}f\,\mathrm{d}x$$

for $y, w \in Y$ and $y = (u, p, a, b)$ and $w = (v, q, e, f)$ is coercive on the domain of ψ.

Proof Young's inequality and the positive definiteness of $\mathbb{C}$ imply that

$$\|\mathbb{C}^{1/2}\big(\varepsilon(u)-p\big)\|^2 \geq (1-\delta^{-1})c_{\mathbb{C}}\|\varepsilon(u)\|^2 + (1-\delta)c_{\mathbb{C}}\|p\|^2.$$

Since I_S^* is only finite if $p = b$ and $\sigma_y|p| \leq a$ almost everywhere in Ω, we have

$$\begin{aligned}\|\mathbb{H}_{\text{iso}}a\|^2 + \|\mathbb{H}_{\text{kin}}b\|^2 &\geq c_{\text{iso}}\|a\|^2 + c_{\text{kin}}\|b\|^2\\ &\geq \max\{c_{\text{iso}}/\sigma_y, c_{\text{kin}}\}\|p\|^2 = c_{\mathbb{H}}\|p\|^2.\end{aligned}$$

Upon choosing $\delta = 1 + c_{\mathbb{H}}/(2c_{\mathbb{C}})$ and using Korn's inequality $\|u\|_{H^1(\Omega)} \leq c\|\varepsilon(u)\|$ the combination of the estimates proves the assertion. □

Remarks 11.1 (i) More generally, it suffices to assume that $\mathbb{H}$ is positive definite and that $\|p\| \leq c\|\xi\|$ on the domain of ψ to guarantee that $\mathscr{A}$ is coercive on dom ψ.
(ii) For kinematic hardening, we identify $p = b$ and then have that $\mathscr{A}$ is coercive on the entire space $Y = H^1_{\mathrm{D}}(\Omega;\mathbb{R}^d) \times L^2(\Omega;\mathbb{R}^{d\times d}_{\text{sym}})$.
(iii) Coercivity does not hold in the case of perfect plasticity when $\mathbb{H}_{\text{iso}} = 0$ and $\mathbb{H}_{\text{kin}} = 0$.
(iv) The von Mises yield criterion $|\operatorname{dev}(\sigma+\beta)| \leq \sigma_y(1+\alpha_+)$ is also called J_2-plasticity since it is based on the second deviatoric stress invariant.
(v) The Tresca yield criterion is based on the maximum shear stress $\sigma_{shear} = \max_{1\leq i,j\leq d}|\sigma_i - \sigma_j|$ with the principal stresses $\sigma_1, \sigma_2, \ldots, \sigma_d$.

The functional $\psi(w) = \int_\Omega I_S^*(q,\zeta)\,\mathrm{d}x$ for $w = (v, q, \zeta)$ is homogeneous of degree one, i.e., we have

$$\psi(\gamma w) = \gamma\psi(w)$$

for all $w \in Y$ and $\gamma \geq 0$. This property has important implications that can be verified by straightforward computations.

Lemma 11.2 (Degree-one homogeneity) *Let $\psi : Y \to \mathbb{R}\cup\{+\infty\}$ be convex, proper, lower semicontinuous, and homogeneous of degree one.*

(i) *With $C_* = \partial\psi(0)$ we have $\partial\psi(w) \subset C_*$ for all $w \in Y$, $0 \in C_*$, and*

$$\psi = I^*_{C_*}.$$

(ii) *For all $w \in Y$ such that $\partial\psi(w) \neq \emptyset$, we have $\langle s, w\rangle = \psi(w)$ for all $s \in \partial\psi(w)$.*

11.1.5 Equivalent Formulations and Existence

In the mathematical description of plastic material behavior, inertial terms were neglected in the equilibrium equation leading to a quasi-stationary evolution problem. Practically, this means that the time-scale of the considered experiment is significantly larger than the internal time scales of a particular material. Mathematically, this induces a *rate-independence* of the problem in the sense that if $y : [0, T] \to Y$ solves the problem subject to the load $\ell : [0, T] \to Y'$, and $\gamma : [0, T'] \to [0, T]$ is an increasing reparametrization of the time interval, then $y \circ \gamma : [0, T'] \to Y$ solves the problem defined by the load $\ell \circ \gamma$. In the elastoplastic model problem this is satisfied since the functional ψ is homogeneous of degree one. This particular property of the problem allows for different notions of solutions. We discuss them in an abstract framework and consider a Hilbert space Y, a continuous and symmetric bilinear form $\mathscr{A} : Y \times Y \to \mathbb{R}$, a function $\ell \in W^{1,\infty}([0, T]; Y')$, and we define the energy functional

$$\mathscr{E}(t, y) = \frac{1}{2}\mathscr{A}(y, y) - \langle \ell(t), y\rangle.$$

A dissipation functional is defined by a proper, convex, lower semicontinuous functional $\psi : Y \to \mathbb{R} \cup \{+\infty\}$ that is degree-one homogeneous. We assume that $\mathscr{A}$ is coercive on dom ψ and let $y_0 \in Y$ be some initial data.

Definition 11.1 The *(primal) evolutionary variational inequality* or *primal problem* seeks $y : [0, T] \to Y$ such that $y(0) = y_0$ and

$$\mathscr{A}\big(y(t), w - \dot{y}(t)\big) - \langle \ell(t), w - \dot{y}(t)\rangle + \psi(w) - \psi\big(\dot{y}(t)\big) \geq 0$$

for all $w \in Y$ and $t \in [0, T]$.

Associating the operator $A : Y \to Y'$ to the bilinear form $\mathscr{A}$, the evolutionary variational inequality is equivalent to the inclusion

$$-Ay + \ell \in \partial\psi\big(\dot{y}\big).$$

The degree-one homogeneity of ψ implies that $\psi = I_{C_*}^*$ with the indicator functional I_{C_*} of the set $C_* = \partial\psi(0)$. Convex duality relations thus yield that we have the equivalent inclusion

$$\dot{y} \in \partial I_{C_*}(-Ay + \ell).$$

Setting $\Sigma = \ell - Ay$ and noting $\dot{y} = A^{-1}(\dot{\ell} - \dot{\Sigma})$ lead to the following formulation.

Definition 11.2 The *dual evolutionary variational inequality* or *dual problem* seeks $\Sigma : [0, T] \to C_*$ such that $\Sigma(0) = \ell(0) - Ay_0$ and

$$\langle \Sigma - \Upsilon, A^{-1}(\dot{\Sigma} - \dot{\ell})\rangle \geq 0$$

for all $\Upsilon \in C_*$ and all $t \in [0, T]$.

Choosing $w = \alpha \widehat{w}$ in the primal problem and considering the limit $\alpha \to \infty$ shows that we have

$$\mathscr{A}(y, \widehat{w}) + \psi(\widehat{w}) \geq \langle \ell, \widehat{w} \rangle.$$

The choice $w = 0$ yields

$$\mathscr{A}(y, \dot{y}) + \psi(\dot{y}) \leq \langle \ell, \dot{y} \rangle.$$

The first inequality implies that, for every $t \in [0, T]$, the element $y(t)$ is a global minimizer for the mapping

$$\widehat{y} \mapsto \mathscr{E}(t, \widehat{y}) + \psi\big(\widehat{y} - y(t)\big).$$

The choice $\widehat{w} = \dot{y}(t)$ and a combination of the inequalities leads to the identity

$$\mathscr{A}(y, \dot{y}) = \langle \ell, \dot{y} \rangle + \psi(\dot{y}).$$

Therefore, we have

$$\frac{d}{dt}\mathscr{E}\big(t, y\big) = \partial_t \mathscr{E}\big(t, y\big) + \langle \partial_y \mathscr{E}\big(t, y\big), \dot{y} \rangle = \partial_t \mathscr{E}\big(t, y\big) - \psi\big(\dot{y}\big).$$

These observations justify the third definition of a solution for the evolution problem.

Definition 11.3 The *energetic formulation* seeks $y : [0, T] \to Y$ such that $y(0) = y_0$ and the *global stability* and *global energy balance* equations

$$\begin{aligned}
&\mathscr{E}\big(t, y(t)\big) \leq \mathscr{E}\big(t, \widehat{y}(t)\big) + \psi\big(\widehat{y} - y(t)\big), \\
&\mathscr{E}\big(t, y(t)\big) + \int_0^t \psi\big(\dot{y}(s)\big)\, \mathrm{d}s = \mathscr{E}\big(0, y(0)\big) - \int_0^t \langle \dot{\ell}(s), y(s) \rangle\, \mathrm{d}s
\end{aligned}$$

hold for all $\widehat{y} \in Y$ and all $t \in [0, T]$.

An advantage of the energetic formulation is that no derivatives of $\mathscr{E}$ or ψ are involved. The global energy balance states that dissipated energy in the time interval $[0, t]$ equals the difference of the change in the stored energy and the power of external forces. Solutions for rate-independent evolution problems can be constructed by an implicit discretization in time, which leads to incremental minimization problems defined by the functionals

$$\widehat{y} \mapsto I_\tau^k(\widehat{y}) = \mathscr{E}(t_k, \widehat{y}) + \psi(\widehat{y} - y^{k-1}).$$

By establishing appropriate a priori bounds and carrying out a passage to a limit, one can prove the following theorem.

Theorem 11.1 (Existence and uniqueness) *If* $\ell \in W^{1,\infty}([0,T];Y')$, $\mathscr{A}$ *is coercive on* $\operatorname{dom}\psi$, *and* $\ell(0) - Ay_0 \in C_*$, *then the energetic formulation and the primal problem have a unique solution* $y \in W^{1,\infty}([0,T];Y)$.

Proof (*sketched*) We recall that the variational inequality is equivalent to the inclusion $\dot{y} \in \partial I_{C_*}(-Ay+\ell)$ and, assuming for simplicity that $\mathscr{A}$ is coercive on the entire space, we introduce the variable $z = y - A^{-1}\ell$. Then $\dot{z} \in \partial I_{C_*}^*(-Az) + A^{-1}\dot{\ell}$. The operator $v \mapsto \partial I_{C_*}(-Av)$ is maximally monotone and Theorem 2.8 implies the existence of a unique solution provided $\partial I_{C_*}(-Az_0) \neq \emptyset$, i.e., $-Ay_0 + \ell(0) \in C_*$. □

A stability and uniqueness result follows from the primal formulation.

Remark 11.2 Let $y_1, y_2 \in W^{1,\infty}([0,T];Y)$ be solutions subject to the loads $\ell_1, \ell_2 \in W^{1,\infty}([0,T];Y')$, respectively. We then have $\alpha \|\dot{y}_j\|_{L^\infty([0,T];Y)} \le \Lambda_j$, $j = 1, 2$, and

$$\alpha^2 \|y_1 - y_2\|^2_{L^\infty([0,T];Y)} \le \alpha \|y_1(0) - y_2(0)\|^2_{\mathscr{A}} + (\Lambda_1 + \Lambda_2) \int_0^T \|\ell_1 - \ell_2\|_{Y'} \,\mathrm{d}t,$$

where α is the coercivity constant of $\mathscr{A}$, $\|y\|^2_{\mathscr{A}} = \mathscr{A}(y,y)$, and $\Lambda_j = \|\dot{\ell}_j\|_{L^\infty([0,T];Y')}$ for $j = 1, 2$.

The theorem implies the existence of a unique solution of the primal problem in the case of positive hardening. Existence of solutions for perfect plasticity and for the dual formulation require additional assumptions.

Remarks 11.3 (i) Although the dual problem is formally equivalent to the primal problem, existence theories require imposing a safe-load assumption, i.e., that there exists a regular stress in the elastic domain that compensates the given loads. The assumption can be proved for individual cases of isotropic and kinematic hardening, cf. [6].
(ii) The existence of solutions for perfect plasticity can be established under a suitable safe-load assumption and within the space of bounded deformations $BD(\Omega)$, i.e., deformations $u \in L^1(\Omega;\mathbb{R}^d)$ such that the symmetric part $\varepsilon(u)$ of the distributional gradient Du is a bounded Radon measure. The solutions can be obtained as vanishing hardening limits, cf. [3, 4, 6].

11.2 Approximation of Rate-Independent Evolutions

For a Hilbert space Y, a symmetric, continuous bilinear form $\mathscr{A} : Y \times Y \to \mathbb{R}$, a convex, proper, lower-semicontinuous functional $\psi : Y \to \mathbb{R} \cup \{+\infty\}$, and a function $\ell \in W^{2,\infty}([0,T];Y')$, we consider the evolution problem

$$\mathscr{A}(y, v - \dot{y}) - \langle \ell(t), v - \dot{y} \rangle + \psi(v) - \psi(\dot{y}) \ge 0$$

for all $v \in Y$, $t \in [0, T]$, subject to the initial condition $y(0) = y_0$. We assume that ψ is homogeneous of degree one and $\mathscr{A}$ is coercive on Y so that there exists a unique solution $y \in W^{1,\infty}([0, T]; Y)$. With the bilinear form $\mathscr{A}$, we associate the invertible, bounded linear operator $A : Y \to Y'$. The formulation is then equivalent to the inclusion $-Ay + \ell \in \partial\psi(\dot{y})$. The norm induced by $\mathscr{A}$ is denoted by $\|\cdot\|_{\mathscr{A}}$ and the norm in Y by $\|\cdot\|$. We follow ideas from [2, 7].

11.2.1 Time-Incremental Minimization

An implicit discretization of the evolution problem can be formulated as a sequence of minimization problems.

Algorithm 11.1 (*Implicit discretization*) Given $y^0 \in Y$ and $\tau > 0$, set $t_k = k\tau$, $k = 0, 1, \ldots, K$, and let $(y^k)_{k=1,\ldots,K} \subset Y$ be a sequence of minimizers for the functionals

$$I_\tau^k(w) = \psi(w - y^{k-1}) + \frac{1}{2}\mathscr{A}(w, w) - \langle \ell(t_k), w\rangle.$$

The iterates of the algorithm are uniquely defined.

Proposition 11.2 (Existence of semi-discrete iterates) *For $k = 1, 2, \ldots, K$, there exists a unique minimizer $y^k \in Y$ for I_τ^k, and we have*

$$\mathscr{A}(y^k, v - d_t y^k) - \langle \ell(t_k), v - d_t y^k\rangle + \psi(v) - \psi(d_t y^k) \geq 0$$

for all $v \in Y$. In particular, $-Ay^k + \ell(t_k) \in C_*$ for $k = 1, 2, \ldots, K$.

Proof The existence of a minimizer in every time step follows from the direct method in the calculus of variations, and we have $0 \in \partial I_\tau^k(y^k)$, i.e.,

$$0 \in Ay^k - \ell(t_k) + \partial\psi(y^k - y^{k-1})$$

and this implies that $-Ay^k + \ell(t_k) \in C_*$ and the variational inequality by incorporating the degree-one homogeneity of ψ. Uniqueness follows from the coercivity of $\mathscr{A}$. □

Noting that $\psi = I_{C_*}^*$ for $C_* = \partial\psi(0)$, the transformation $z = -y + A^{-1}\ell$ shows that the evolution problem is equivalent to the inclusion $-\dot{z} \in \partial I_{C_*}(Az) - A^{-1}\dot{\ell}$. Similarly, the transformation $z^k = -y^k + A^{-1}\ell^k$ shows that the variational inequality of the proposition is equivalent to the inclusion $-d_t z^k \in \partial I_{C_*}(Az^k) - A^{-1}d_t\ell^k$. We abbreviate $r = A^{-1}\dot{\ell}$ and $r^k = A^{-1}d_t\ell(t_k)$ in the following and apply the abstract strategy of Theorem 4.7 to the rate-independent evolution problem.

Theorem 11.2 (Auxiliary error estimate) *Suppose that* $z \in W^{1,\infty}([0,T];Y)$ *satisfies*

$$-\dot{z} + r \in \partial I_{C_*}(Az)$$

and the sequence $(z^k)_{k=0,\dots,K} \subset Y$ *is such that* $z(0) = z^0$, $Az^k \in C_*$ *for* $k = 0, 1, \dots, K$, *and*

$$-d_t z^k + r^k \in \partial I_{C_*}(Az^k)$$

for $k = 1, 2, \dots, K$. *With the piecewise affine interpolant* $\widehat{z}_\tau : [0,T] \to Y$ *of the approximations* $(z^k)_{k=0,\dots,K}$, *we have*

$$\sup_{t\in[0,T]} \|z - \widehat{z}_\tau\|_{\mathscr{A}} \le \tau\big(\|r^0\|_{\mathscr{A}} + cT\|r\|_{W^{1,\infty}([0,T];Y)}\big).$$

Proof (i) We first note that the discrete inclusion is equivalent to the variational inequality

$$\langle -d_t z^k + r^k, v - Az^k\rangle \le 0$$

for all $v \in C_*$. With the choice $v = Az^{k-1}$ we define

$$-\mathscr{E}_k = \tau\|d_t z^k\|_{\mathscr{A}}^2 - \tau\langle r^k, A d_t z^k\rangle \le 0$$

and note that

$$\|d_t z^k\|_{\mathscr{A}} \le \|r^k\|_{\mathscr{A}}.$$

(ii) We let $z_\tau^+ : [0,T] \to Y$ be the piecewise constant function satisfying $z_\tau^+(t) = z^k$ if $t_{k-1} < t \le t_k$. Similarly, $r_\tau^+ : [0,T] \to Y'$ denotes the piecewise constant interpolant of $(r^k)_{k=1,\dots,K}$. We have

$$\langle -\partial_t \widehat{z}_\tau + r^+, v - Az_\tau^+\rangle \le 0$$

for all $v \in C_*$. Defining

$$\mathscr{C}_\tau(t) = \langle -\partial_t \widehat{z}_\tau + r^+, Az_\tau^+ - A\widehat{z}_\tau\rangle,$$

we find

$$\langle -\partial_t \widehat{z}_\tau + r, v - A\widehat{z}_\tau\rangle \le \mathscr{C}_\tau(t) + \langle r - r_\tau^+, v - A\widehat{z}_\tau\rangle.$$

(iii) Noting the equation for z, i.e.,

$$\langle -\partial_t z + r, v - Az\rangle \le 0$$

and choosing $v = Az$ and $v = A\widehat{z}_\tau$ in the equations for $\widehat{z}_\tau$ and z, respectively, and adding the inequalities, we find that

$$\frac{1}{2}\frac{d}{dt}\|z - \widehat{z}_\tau\|_{\mathscr{A}}^2 \le \mathscr{C}_\tau(t) + \frac{T}{2}\|r - r_\tau^+\|_{\mathscr{A}}^2 + \frac{1}{2T}\sup_{t\in[0,T]}\|z - \widehat{z}_\tau\|_{\mathscr{A}}^2.$$

Using that $z_\tau^+ - \widehat{z}_\tau = -(t-t_k)d_t z^k$ for $t \in [t_{k-1}, t_k]$ we have

$$\mathscr{C}_\tau(t) = (t-t_k)\|d_t z^k\|_{\mathscr{A}}^2 - (t-t_k)\langle r^k, Ad_t z^k\rangle = \frac{t-t_k}{\tau}(-\mathscr{E}_k) \le \mathscr{E}_k.$$

We also note that $\|r - r_\tau^+\|_{\mathscr{A}} \le \tau \sup_{t\in[0,T]} \|\dot{r}\|_{\mathscr{A}}$. An integration over the interval $[0, t^*]$, where $t^* \in [0,T]$ corresponds to the maximum of $t \mapsto \|z - \widehat{z}_\tau\|_{\mathscr{A}}$, and $z(0) = z^0$ show that

$$\frac{1}{2}\sup_{t\in[0,T]} \|z - \widehat{z}_\tau\|_{\mathscr{A}}^2 \le \tau \sum_{k=1}^K \mathscr{E}_k + \tau^2 \frac{T^2}{2} \sup_{t\in[0,T]} \|\dot{r}\|_{\mathscr{A}}^2.$$

(iv) We note that the equation for z^{k-1} with $k \ge 2$ reads as

$$\langle -d_t z^{k-1} + r^{k-1}, v - Az^{k-1}\rangle \le 0.$$

By defining z^{-1}, so that $-d_t z^0 + r^0 = 0$, and noting that $0 \in \partial^0 I_{C_*}(Az^0)$, this variational inequality also holds with $k = 1$. The choice $v = Az^k$ yields

$$\langle -d_t z^{k-1} + r^{k-1}, Ad_t z^k\rangle \le 0.$$

With the definition of $\mathscr{E}_k$ we deduce that

$$\begin{aligned}
\mathscr{E}_k &= -\tau\langle d_t z^k + r_k, Ad_t z^k\rangle \\
&\le -\tau\langle d_t z^k - r_k, Ad_t z^k\rangle + \tau\langle d_t z^{k-1} - r^{k-1}, Ad_t z^k\rangle \\
&= -\tau^2\langle d_t^2 z^k - d_t r^k, Ad_t z^k\rangle \\
&= -\tau^2 \frac{d_t}{2}\|d_t z^k\|_{\mathscr{A}}^2 - \frac{\tau^3}{2}\|d_t^2 z^k\|_{\mathscr{A}}^2 + \tau^2\langle d_t r^k, Ad_t z^k\rangle \\
&\le -\tau^2 \frac{d_t}{2}\|d_t z^k\|_{\mathscr{A}}^2 + \tau^2\|d_t r^k\|_{\mathscr{A}}\|d_t z^k\|_{\mathscr{A}} \\
&\le -\tau^2 \frac{d_t}{2}\|d_t z^k\|_{\mathscr{A}}^2 + \tau^2 \sup_{t\in[0,T]} \|\dot{r}\|_{\mathscr{A}}\|r\|_{\mathscr{A}},
\end{aligned}$$

where we used $\|d_t z^k\|_{\mathscr{A}} \le \|r^k\|_{\mathscr{A}}$ and $\|d_t r^k\|_{\mathscr{A}} \le \sup_{t\in[0,T]} \|\dot{r}\|_{\mathscr{A}}$. A summation of $\mathscr{E}_k$ over $k = 1, 2, \dots, K$ yields

$$\tau \sum_{k=1}^K \mathscr{E}_k \le \frac{\tau^2}{2}\|d_t z^0\|_{\mathscr{A}}^2 + c\tau^2 T\|r\|_{W^{1,\infty}([0,T];Y')}^2.$$

This implies the assertion. □

Remark 11.4 The proof of the theorem provides the computable a posteriori error estimate

$$\sup_{t\in[0,T]} \|z-\widehat{z}\|_{\mathscr{A}}^2 \le 2\tau \sum_{k=1}^K \mathscr{E}_k + \tau^3 \sum_{k=1}^K \sup_{t\in[t_{k-1},t_k]} \|\dot{r}\|_{\mathscr{A}}^2.$$

The theorem implies an error estimate for the approximation of the original formulation.

Corollary 11.1 (Time discretization) *Assume that* $y_0 \in Y$ *satisfies* $-Ay_0 + \ell(0) \in \partial\psi(0)$ *and that we have* $\ell \in W^{2,\infty}([0,T];Y')$. *For the solution* $y \in W^{1,\infty}([0,T];Y)$ *of the evolutionary variational inequality with* $y(0) = y_0$ *and the piecewise affine interpolant* $\widehat{y}_\tau \in W^{1,\infty}([0,T];Y)$ *of the iterates* $(y^k)_{k=0,\dots,K}$ *of Algorithm* 11.1 *such that* $y^0 = y_0$, *we have*

$$\sup_{t\in[0,T]} \|y-\widehat{y}_\tau\| \le c\tau.$$

Example 11.3 The error estimate applies to kinematic hardening. Recalling that in this case we have with the identification $p = b$ and thus $\beta = -\mathbb{H}_{\rm kin} p$ that

$$\mathscr{A}(y,w) = \int_\Omega \mathbb{C}(\varepsilon(u)-p) : (\varepsilon(v)-q) + \mathbb{H}_{\rm kin} p : q \,{\rm d}x$$

for $y = (u,p)$ and $w = (v,q)$. Hence, with $\sigma = \mathbb{C}(\varepsilon(u)-p)$ we have that $Ay = (-\operatorname{div}\sigma, -\sigma + \mathbb{H}_{\rm kin} p)$. We recall that $\psi(y) = \sigma_y |p|_0$, where $|p|_0 = |p|$ if $\operatorname{tr} p = 0$ and $|p|_0 = +\infty$ otherwise, and

$$\ell(t,w) = \int_\Omega f(t)\cdot v \,{\rm d}x + \int_{\Gamma_{\rm N}} g(t)\cdot v \,{\rm d}s.$$

With $\sigma_0 = \sigma(0)$ the compatibility condition $-Ay_0 + \ell(0) \in \partial\psi(0)$ is thus equivalent to $\operatorname{div}\sigma_0 + f(0) = 0$ in Ω, $\sigma_0 n = g$ on $\Gamma_{\rm N}$, and $\sigma_0 - \mathbb{H}_{\rm kin} p_0 \in \sigma_y \partial|\cdot|_0$, that is, $|\operatorname{dev}(\sigma_0 - \mathbb{H}_{\rm kin} p_0)| = |\operatorname{dev}(\sigma_0 + \beta_0)| \le \sigma_y$.

Remark 11.5 The error estimate also holds for a spatially discrete version of the problem. In this case Y is replaced by a finite-dimensional subspace $Y_h \subset Y$ and the subdifferential is defined with respect to this space.

11.2.2 Discretization in Space

We next investigate the error introduced by a spatial discretization. For this we assume that we are given a finite-dimensional subspace $Y_h \subset Y$ and let $P_{\mathscr{A},h} : Y \to Y_h$

denote the orthogonal projection onto Y_h with respect to $\mathscr{A}$; that is, for $z \in Y$ we let $P_{\mathscr{A},h} z \in Y_h$ be such that

$$\mathscr{A}(P_{\mathscr{A},h} z - z, v_h) = 0$$

for all $v_h \in Y_h$. We assume that there exists a bounded linear operator $\mathscr{J}_h : Y \to Y_h$ such that $\mathscr{J}_h z \in \operatorname{dom} \psi$ whenever $z \in \operatorname{dom} \psi$.

Proposition 11.3 (Space discretization) *Let $y \in W^{1,\infty}([0,T];Y)$ be the solution of the primal problem and let $y_h \in W^{1,\infty}([0,T];Y)$ be the uniquely defined function $y_h : [0,T] \to Y_h$ satisfying $y_h(0) = y_h^0 = P_{\mathscr{A},h} y_0$ and*

$$\mathscr{A}(y_h, v_h - \dot{y}_h) - \langle \ell(t), v_h - \dot{y}_h \rangle + \psi(v_h) - \psi(\dot{y}_h) \geq 0$$

for all $v_h \in Y_h$ and $t \in [0,T]$. Assume that there exists c_ψ such that

$$|\psi(v) - \psi(w)| \leq c_\psi \|v - w\|$$

for all $v, w \in$ dom ψ. We then have

$$\sup_{t \in [0,T]} \|y - y_h\|^2 \leq c \int_0^T \|(1 - \mathscr{J}_h)\dot{y}\| \, \mathrm{d}t + \|(1 - P_{\mathscr{A},h}) y_0\|^2.$$

Proof The existence of the spatially discrete solution follows as in the continuous case. We choose $v = \dot{y}_h$ and add the discrete formulation to the continuous variational inequality to verify that

$$\mathscr{A}\big(y_h, \dot{y} - \dot{y}_h) + \mathscr{A}(y, \dot{y}_h - \dot{y}) + \mathscr{A}(y_h, v_h - \dot{y}) - \langle \ell(t), v_h - \dot{y} \rangle + \psi(v_h) - \psi(\dot{y}) \geq 0.$$

The choice $v_h = \mathscr{J}_h \dot{y}$ yields

$$\begin{aligned} \mathscr{A}(y_h - y, \dot{y}_h - \dot{y}) &\leq -\langle \ell(t), \mathscr{J}_h \dot{y} - \dot{y} \rangle + \psi(\mathscr{J}_h \dot{y}) - \psi(\dot{y}) + \mathscr{A}(y_h, \mathscr{J}_h \dot{y} - \dot{y}) \\ &\leq \big(\|\ell(t)\| + c_\psi + c_{\mathscr{A}} \|y_h\|\big) \|\dot{y} - \mathscr{J}_h \dot{y}\|. \end{aligned}$$

We thus have

$$\frac{1}{2} \frac{d}{dt} \|y - y_h\|_{\mathscr{A}}^2 \leq \big(\|\ell(t)\| + c_\psi + c_{\mathscr{A}} \|y_h\|\big) \|\dot{y} - \mathscr{J}_h \dot{y}\|$$

which implies the asserted estimate. □

Remarks 11.6 (i) For a sequence of dense subspaces $(Y_h)_{h>0}$, the estimate of the proposition implies the convergence $y_h \to y$ as $h \to 0$ provided $\mathscr{J}_h$ has appropriate

approximation properties. Related convergence rates under regularity assumptions on $\dot{y}$ are discussed in Example 11.4.
(ii) For kinematic hardening, the condition on the operator $\mathscr{J}_h$ means that for $(\widetilde{u}_h, \widetilde{p}_h) = \mathscr{J}_h(u, p)$ we have $\operatorname{tr} \widetilde{p}_h = 0$ if $\operatorname{tr} p = 0$. This can be guaranteed by employing averaging operators for the definition of $\widetilde{p}_h$. The proof of the proposition simplifies if we can choose $\mathscr{J}_h = P_{\mathscr{A},h}$.

11.2.3 Fully Discrete Approximation

The combination of the estimates for the semi-discrete schemes allows us to derive an error estimate for fully discrete approximations. These are obtained with the following algorithm.

Algorithm 11.2 (*Fully discrete iteration*) Given $y_h^0 \in Y_h$ and $\tau > 0$, let $(y_h^k)_{k=1,\dots,K}$ be a sequence of minimizers $y_h^k \in Y_h$ for the functionals

$$I_{\tau,h}^k(w_h) = \psi\big(w_h - y_h^{k-1}\big) + \frac{1}{2}\mathscr{A}(w_h, w_h) - \langle \ell(t_k), w_h \rangle.$$

The iterates of the algorithm are uniquely defined and satisfy a discrete variational inequality.

Proposition 11.4 (Existence of fully discrete approximations) *There exists a unique discrete solution $(y_h^k)_{k=0,\dots,K}$ and we have*

$$\mathscr{A}(y_h^k, v_h - d_t y_h^k) - \langle \ell(t_k), v_h - d_t y_h^k \rangle + \psi(v_h) - \psi(d_t y_h^k) \ge 0$$

for $k = 1, 2, \dots, K$ and all $v_h \in Y_h$. If $-Ay_0 + \ell(0) \in \partial\psi(0)$ and $y_h^0 = P_{\mathscr{A},h} y_0$, then we have

$$\max_{k=1,\dots,K} \|d_t y_h^k\| \le c\|\ell\|_{W^{1,\infty}([0,T];Y')}.$$

Proof The derivation of the variational inequality is analogous to the proof of Proposition 11.2. The assumption on y_0 and the definition of y_h^0 imply that

$$-\mathscr{A}(y_h^0, v_h) + \langle \ell(0), v_h \rangle = -\mathscr{A}(y_0, v_h) + \langle \ell(0), v_h \rangle \le \psi(v_h)$$

and by setting $y_h^{-1} = y_h^0$, i.e., $d_t y_h^0 = 0$ the variational inequality also holds for $k = 0$. To prove the estimate we note that the choice $v_h = 0$ yields

$$\mathscr{A}(y_h^k, d_t y_h^k) \le \langle \ell(t_k), d_t y_h^k \rangle - \psi(d_t y_h^k),$$

while the choice $v_h = d_t y_h^k + d_t y_h^{k-1}$ in the equation for y_h^{k-1} leads to

$$-\mathscr{A}(y_h^{k-1}, d_t y_h^k) \le -\langle \ell(t_{k-1}), d_t y_h^k \rangle + \psi(d_t y_h^k + d_t y_h^{k-1}) - \psi(d_t y_h^{k-1}).$$

Adding the two inequalities shows that

$$\begin{aligned}\tau\mathscr{A}(d_t y_h^k, d_t y_h^k) &\le \langle \tau d_t \ell(t_k), d_t y_h^k\rangle + \psi(d_t y_h^k + d_t y_h^{k-1}) - \psi(d_t y_h^k) - \psi(d_t y_h^{k-1})\\ &\le \tau \sup_{t\in[t_k,t_{k-1}]} \|\dot{\ell}\|_{Y'}\|d_t y_h^k\|,\end{aligned}$$

where we used the convexity and degree-one homogeneity of ψ. □

Given the sequence of approximations $(y_h^k)_{k=0,\dots,K}$, we let $\widehat{y}_{h,\tau} \in W^{1,\infty}([0,T];Y)$ denote its piecewise affine interpolant in time. The combination of the estimates for the semi-discrete schemes implies the following error estimate.

Theorem 11.3 (Fully discrete approximations) *Given the sequence of approximations $(y_h^k)_{k=0,\dots,K} \in Y_h$ such that $y_h^0 = P_{\mathscr{A},h}y_0$, we have for its piecewise affine interpolant $\widehat{y}_{h,\tau}$ that*

$$\sup_{t\in[0,T]} \|y - \widehat{y}_{h,\tau}\|^2 \le c\big(\tau^2 + \int_0^T \|(1-\mathscr{I}_h)\dot{y}\|\,\mathrm{d}t + \|(1-P_{\mathscr{A},h})y_0\|^2\big).$$

Proof We let $y_h \in W^{1,\infty}([0,T];Y)$ be the solution of the semi-discrete approximation in space, and notice that according to Proposition 11.3 we have

$$\sup_{t\in[0,T]} \|y - y_h\|^2 \le c\int_0^T \|(1-\mathscr{I}_h)\dot{y}\|\,\mathrm{d}t + \|(1-P_{\mathscr{A},h})y_0\|^2.$$

The fully discrete scheme is interpreted as a temporal discretization of the semi-discrete scheme in space and the arguments of the proof of Theorem 11.2 lead to the estimate

$$\sup_{t\in[0,T]} \|y_h - \widehat{y}_{h,\tau}\| \le c\tau.$$

The combination of the two estimates implies the estimate of the theorem. □

Examples 11.4 (i) For kinematic hardening, we may eliminate the internal variables and have $Y = H^1_{\mathrm{D}}(\Omega;\mathbb{R}^d) \times L^2(\Omega;\mathbb{R}^{d\times d})$ and $y = (u,p) \in Y$. Assuming that $u \in W^{1,1}([0,T];H^2(\Omega;\mathbb{R}^d))$ and $p \in W^{1,1}([0,T];H^1(\Omega;\mathbb{R}^{d\times d}))$, we obtain with the subspace $Y_h = \mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h)^d \times \mathscr{L}^0(\mathscr{T}_h)^{d\times d}$, the convergence rate $\mathscr{O}(\tau + h^{1/2})$. More realistic regularity assumptions suggest the convergence rate $\mathscr{O}(\tau + h^{1/4-\delta})$ for arbitrary $\delta > 0$.
(ii) If ℓ and ψ are of lower order, e.g., $\ell \in W^{2,\infty}([0,T];L^2(\Omega;\mathbb{R}^d)')$ and $\psi : L^2(\Omega;\mathbb{R}^{d\times d}) \to \mathbb{R}\cup\{+\infty\}$, and Y is elliptic on $H^1_{\mathrm{D}}(\Omega;\mathbb{R}^d) \times H^1(\Omega;\mathbb{R}^{d\times d})$, e.g., in the case of gradient plasticity, then no regularity is required to deduce the convergence rate $\mathscr{O}(\tau + h^{1/2})$ for lowest order conforming finite elements.

11.2.4 A Posteriori Error Control

A basis for full a posteriori error control is a characterization of the solution of the evolution problem as the unique minimizer of an appropriate functional. The key ingredient for this is the Fenchel duality relation

$$\psi(y) + \psi^*(y') \geq \langle y, y' \rangle$$

in which equality holds if and only if $y \in \partial\psi^*(y')$ and $y' \in \partial\psi(y)$.

Proposition 11.5 (Minimization property [10]) *The function $y \in W^{1,1}([0,T];Y)$ satisfies*

$$-Ay + \ell \in \partial\psi(\dot{y}), \quad y(0) = y_0$$

if and only if $F(y) = 0$ for the nonnegative functional

$$F(z) = \int_0^T \psi(\dot{z}) + \psi^*(\ell - Az) - \langle \ell - Az, \dot{z} \rangle \,\mathrm{d}t + \chi\big(z(0) - y_0\big)$$

defined for $z \in W^{1,1}([0,T];Y)$ and with $\chi(w) = (1/2)\langle Aw, w\rangle + \|w\|^2$.

Proof The Fenchel duality relation implies that $F(z) \geq 0$. If $F(y) = 0$, then it also implies that $-Ay + \ell \in \partial\psi(\dot{y})$ and $y(0) = y_0$ is an immediate consequence. The converse implication follows analogously. □

Theorem 11.4 (A posteriori error control [10]) *For $y \in W^{1,1}([0,T];Y)$ with $F(y) = 0$ and every $v \in W^{1,1}([0,T];Y)$, we have*

$$\frac{1}{4} \sup_{t\in[0,T]} \|y - v\|_{\mathscr{A}}^2 \leq F(v).$$

Proof For $s \in [0,T]$ we define

$$F^s(z) = \int_0^s \psi(\dot{z}) + \psi^*(\ell - Az) - \langle \ell - Az, \dot{z} \rangle \,\mathrm{d}t + \chi\big(z(0) - y_0\big).$$

Noting that $(d/dt)\|z\|_{\mathscr{A}}^2 = 2\langle Az, \dot{z}\rangle$ and incorporating the definition of χ implies that

$$\begin{aligned} F^s(z) = &\int_0^s \psi(\dot{z}) + \psi^*(\ell - Az) - \langle \ell, \dot{z} \rangle \,\mathrm{d}t \\ &+ \frac{1}{2}\big(\|z(s)\|_{\mathscr{A}}^2 - \|z(0)\|_{\mathscr{A}}^2\big) + \chi\big(z(0) - y_0\big) \end{aligned}$$

$$
\begin{aligned}
&= \int_0^s \psi(\dot{z}) + \psi^*(\ell - Az) - \langle \ell, \dot{z}\rangle \,\mathrm{d}t \\
&\quad + \frac{1}{2}\big(\|z(s)\|_{\mathscr{A}}^2 + \|y_0\|_{\mathscr{A}}^2\big) - \langle Az(0), y_0\rangle + \|z(0) - y_0\|^2.
\end{aligned}
$$

This shows that $G^s(z) = F^s(z) - \|z(s)\|_{\mathscr{A}}^2/2$ is convex and hence F^s is the sum of a convex and a quadratic function. For $\theta \in [0,1]$ and $v, w \in W^{1,1}([0,T]; Y)$ and $z = \theta v + (1-\theta)w$, we thus deduce that

$$
\begin{aligned}
0 \le F^s(z) &= G^s(z) + \frac{1}{2}\|z(s)\|_{\mathscr{A}}^2 \\
&\le \theta G^s(v) + (1-\theta)G^s(w) + \frac{1}{2}\|\theta v(s) + (1-\theta)w(s)\|_{\mathscr{A}}^2.
\end{aligned}
$$

Incorporating the formula

$$
\theta\phi(v) + (1-\theta)\phi(w) - \phi(\theta v + (1-\theta)w) = \theta(1-\theta)\phi(v-w)
$$

for $\phi(v) = \|v\|_{\mathscr{A}}^2/2$ implies the estimate

$$
\frac{\theta(1-\theta)}{2} \sup_{s\in[0,T]} \|v-w\|_{\mathscr{A}}^2 \le \theta F(v) + (1-\theta)F(w).
$$

For $\theta = 1/2$ and $w = y$ we deduce the asserted estimate. □

Example 11.5 Consider kinematic hardening with the variable $y = (u,p)$ and let $\widehat{y} = (\widehat{u}, \widehat{p})$ be an admissible function, i.e., $\psi^*(\ell - A\widehat{y}) = I_{C_*}(\ell - A\widehat{y}) = 0$ for all $t \in [0,T]$. Assume that $\widehat{y}$ is piecewise affine in time with respect to the time steps $(t_k)_{k=0,\dots,K}$ and let $(\widehat{u}^k, \widehat{p}^k) = (\widehat{u}(t_k), \widehat{p}(t_k))$ and $\tau_k = t_k - t_{k-1}$. We then have

$$
\sup_{t\in[0,T]} \|y - \widehat{y}\|_{\mathscr{A}}^2 \le 4\sum_{k=1}^K \tau_k \eta_k^2(\widehat{u}, \widehat{p})
$$

with

$$
\eta_k^2(\widehat{u}, \widehat{p}) = \int_\Omega \sigma_y |d_t\widehat{p}^k| - \overline{f}(t_k)\cdot d_t\widehat{u}^k + \mathbb{C}\big(\varepsilon(\widehat{u}^k) - \widehat{p}^k\big) : \big(\varepsilon(d_t\widehat{u}^k) - d_t\widehat{p}^k\big)\,\mathrm{d}x,
$$

where $\overline{f}(t_k) = \tau_k^{-1}\int_{t_{k-1}}^{t_k} f(t)\,\mathrm{d}t$. The admissible function $\widehat{y}$ can be constructed in a post-processing procedure from a typically inadmissible finite element approximation $\widehat{y}_{h,\tau}$ and the triangle inequality leads to a computable estimate for the approximation error $\|y - \widehat{y}_{h,\tau}\|$. Lowest order approximations satisfy the stress admissibility condition $\sigma_h \in K$ exactly, but in general not the equilibrium condition $-\operatorname{div}\sigma_h = f$.

11.3 Numerical Solution

The numerical solution of the nonlinear system of equations related to a time step in the implicit discretization of quasi-static elastoplastic evolution problems can be formulated as a nonlinear displacement problem. The key to this reformulation is a pointwise solution of the discretized flow rule that leads to a formula for the stress field in terms of the displacement. We employ ideas from [1, 8, 9].

11.3.1 Solution of the Discretized Flow Rule

For a step size $\tau > 0$, an implicit discretization of the flow rule reads as

$$\Sigma^k \in \partial I_S^*(d_t P^k)$$

with the generalized stress $\Sigma^k = (\sigma^k, \chi^k)$, the generalized plastic strain $P^k = (p^k, \xi^k)$, and the backward difference quotient $d_t P^k = (P^k - P^{k-1})/\tau$. We recall that we have the relations

$$\chi^k = -\mathbb{H}\xi^k, \quad \sigma^k = \mathbb{C}(\varepsilon(u^k) - p^k).$$

The following proposition shows that for the generalized plastic strain P^{k-1}, the generalized stress Σ^{k-1} from a previous time step, and a trial strain $\varepsilon(u^k)$ that may be a guess of the true displacement in the k-th time step, the corresponding generalized stress Σ^k is pointwise uniquely determined by the flow rule and defines P^k, cf. Fig. 11.3.

Proposition 11.6 (Return map) *Given arbitrary P^{k-1}, Σ^{k-1}, and $\varepsilon(u^k)$, there exist uniquely defined Σ^k and P^k such that*

$$\Sigma^k \in \partial I_S^*(d_t P^k).$$

The field Σ^k is the best approximation of $\widetilde{\Sigma}^k = (\tau \mathbb{C} d_t \varepsilon(u^k) + \sigma^{k-1}, \chi^{k-1})$ in S with respect to the scalar product $\langle(\sigma, \chi), (p, \zeta)\rangle = (\sigma : \mathbb{C}^{-1} p + \chi : \mathbb{H}^{-1}\zeta)/\tau$.

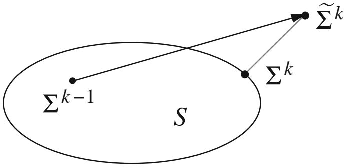


Fig. 11.3 For given P^{k-1}, Σ^{k-1}, and $\varepsilon(u^k)$ the elastic trial stress $\widetilde{\Sigma}^k = (\tau \mathbb{C} d_t \varepsilon(u^k) + \sigma^{k-1}, \chi^{k-1})$ is projected onto the set S to obtain an admissible stress $\Sigma^k \in \partial I_S^*(d_t P^k)$

Proof We note $d_t p^k = d_t\varepsilon(u^k) - \mathbb{C}^{-1} d_t\sigma^k = d_t\varepsilon(u^k) + \tau^{-1}\mathbb{C}^{-1}\sigma^{k-1} - \tau^{-1}\mathbb{C}^{-1}\sigma^k$ and $d_t\xi^k = -\mathbb{H}^{-1} d_t\chi^k = \tau^{-1}\mathbb{H}^{-1}\chi^{k-1} - \tau^{-1}\mathbb{H}^{-1}\chi^k$. Hence, the inclusion $d_t P^k \in \partial I_S(\Sigma^k)$ is equivalent to

$$-\tau^{-1}\begin{bmatrix}\mathbb{C}^{-1} & 0\\ 0 & \mathbb{H}^{-1}\end{bmatrix}\left(\Sigma^k - \begin{bmatrix}\tau\mathbb{C} d_t\varepsilon(u^k) + \sigma^{k-1}\\ \chi^{k-1}\end{bmatrix}\right) \in \partial I_S(\Sigma^k).$$

This implies the assertion. □

The proposition shows that given Σ^{k-1} and u^{k-1}, which uniquely define P^{k-1}, there exists a well-defined map

$$\widehat{\mathscr{P}}^k : u^k \mapsto \Sigma^k,$$

so that the flow rule is satisfied with P^k determined by Σ^k and u^k. For the von Mises yield criterion we derive an explicit formula for the operator $\widehat{\mathscr{P}}^k$. An essential ingredient for this is the following lemma in which we employ the functional

$$|\cdot|_0 : \mathbb{R}^{d\times d}_{\text{sym}} \to \mathbb{R}\cup\{+\infty\}, \quad \dot{p} \mapsto \begin{cases}|\dot{p}| & \text{if } \operatorname{tr}\dot{p} = 0,\\ +\infty & \text{otherwise.}\end{cases}$$

We assume that

$$\mathbb{C}E = \lambda(\operatorname{tr} E) I_d + 2\mu E$$

for $E \in \mathbb{R}^{d\times d}_{\text{sym}}$ and constants $\lambda, \mu > 0$.

Lemma 11.3 (Explicit solution [1]) *Let $B \in \mathbb{R}^{d\times d}_{\text{sym}}$ and $\dot{p} \in \mathbb{R}^{d\times d}_{\text{sym}}$ with* $\operatorname{tr}\dot{p} = 0$*, and $\eta, r \ge 0$ be such that*

$$z = B - \tau(\mathbb{C} + 2\eta)\dot{p} \in \partial r|\cdot|_0(\dot{p}).$$

Then

$$\dot{p} = \frac{(|\operatorname{dev} B| - r)_+}{2\tau(\mu+\eta)}\frac{\operatorname{dev} B}{|\operatorname{dev} B|}.$$

Proof The inclusion is equivalent to $z : (q - \dot{p}) + r|\dot{p}|_0 \le r|q|_0$. If $\dot{p} = 0$, we have $z : q \le r|q|$ and deduce $|\operatorname{dev} z| = |\operatorname{dev} B| \le r$. If $\dot{p} \ne 0$, then $\operatorname{dev} z = r\dot{p}/|\dot{p}|$ and using $\mathbb{C}\dot{p} = 2\mu\dot{p}$, we find that

$$\operatorname{dev} B - 2(\mu+\eta)\dot{p} = r\frac{\dot{p}}{|\dot{p}|}$$

which implies $|\operatorname{dev} B| = 2(\mu+\eta)|\dot{p}| + r$ and $2|\dot{p}| = (|\operatorname{dev} B| - r)/(\mu+\eta)$. Since $\dot{p}$ and $\operatorname{dev} B$ are parallel, we deduce the asserted formula. □

We solve the discretized flow rule explicitly for the yield function $\Phi(\sigma, \alpha, \beta) = |\operatorname{dev}(\sigma + \beta)| - \sigma_y(1 + \alpha_+)$ and the corresponding support functional of the set of admissible stresses

$$I_S^*(\dot{P}) = \begin{cases} \sigma_y|\dot{p}| & \text{if } \operatorname{tr} \dot{p} = 0,\ \dot{b} = \dot{p},\ \sigma_y|\dot{p}| \le -\dot{a}, \\ +\infty & \text{otherwise,} \end{cases}$$

cf. Lemma 11.1. We assume the constitutive relations

$$\alpha = -\mathbb{H}_{\text{iso}} a, \quad \beta = -\mathbb{H}_{\text{kin}} b$$

such that $\mathbb{H}_{\text{kin}}$ is a multiple of the identity.

Proposition 11.7 (General von Mises flow rule) *Assume that* $(p^{k-1}, a^{k-1}, b^{k-1})$ *with* $\operatorname{tr} p^{k-1} = 0$, $b^{k-1} = p^{k-1}$, *and* $a^{k-1} \le 0$ *are given. For arbitrary* u^k, *define* $A^k = \mathbb{C}\big(\varepsilon(u^k) - p^{k-1}\big)$. *Then with*

$$d_t p^k = \frac{\big(|\operatorname{dev} A^k - \mathbb{H}_{\text{kin}} p^{k-1}| - \sigma_y(1 - \mathbb{H}_{\text{iso}} a^{k-1})\big)_+}{2\tau(\mu + \mathbb{H}_{\text{iso}}\sigma_y^2 + \mathbb{H}_{\text{kin}})} \frac{\operatorname{dev} A^k - \mathbb{H}_{\text{kin}} p^{k-1}}{|\operatorname{dev} A^k - \mathbb{H}_{\text{kin}} p^{k-1}|}$$

and $d_t(a^k, b^k) = (-\sigma_y|d_t p^k|, d_t p^k) = -d_t(\mathbb{H}_{\text{iso}}\alpha^k, \mathbb{H}_{\text{kin}}\beta^k)$ *and* $\sigma^k = A^k - \tau\mathbb{C} d_t p^k$ *we have*

$$\Sigma^k \in \partial I_S^*(d_t P^k).$$

Proof We omit the superscript k and abbreviate $\dot{P} = d_t P^k$ and $P' = P^{k-1}$ in the following. The inclusion $\Sigma \in \partial I_S^*(\dot{P})$ states that $\dot{P}$ is a minimizer for the mapping

$$\Gamma : \dot{P} \mapsto I_S^*(\dot{P}) - \Sigma : \dot{P} = I_S^*(\dot{p}, \dot{a}, \dot{b}) - \sigma : \dot{p} - \alpha\dot{a} - \beta : \dot{b}.$$

The identities $(\alpha, \beta) = -(\mathbb{H}_{\text{iso}} a, \mathbb{H}_{\text{kin}} b)$ and $a = a' + \tau\dot{a}$, $b = b' + \tau\dot{b}$ lead to

$$\begin{aligned} \Gamma(\dot{p}, \dot{a}, \dot{b}) &= I_S^*(\dot{p}, \dot{a}, \dot{b}) - \sigma : \dot{p} + \mathbb{H}_{\text{iso}} a\dot{a} + \mathbb{H}_{\text{kin}} b : \dot{b} \\ &= I_S^*(\dot{p}, \dot{a}, \dot{b}) - \sigma : \dot{p} + \mathbb{H}_{\text{iso}} a'\dot{a} + \tau\mathbb{H}_{\text{iso}}\dot{a}^2 + \mathbb{H}_{\text{kin}} b' : \dot{b} + \tau\mathbb{H}_{\text{kin}}\dot{b} : \dot{b}. \end{aligned}$$

We note that I_S^* is finite only if $\dot{p} = \dot{b}$, so that we may eliminate $\dot{b}$ and b', i.e., we may consider minimizing

$$\Gamma'(\dot{p}, \dot{a}) = I_S^*(\dot{p}, \dot{a}, \dot{p}) - \sigma : \dot{p} + \mathbb{H}_{\text{iso}} a'\dot{a} + \tau\mathbb{H}_{\text{iso}}\dot{a}^2 + \mathbb{H}_{\text{kin}} p' : \dot{p} + \tau\mathbb{H}_{\text{kin}}\dot{p} : \dot{p}.$$

Given $\dot{p}$, the functional I_S^* is finite only if $\sigma_y|\dot{p}|_0 \le -\dot{a}$. Noting $a' \le 0$ and $\dot{a} \le 0$, show that given $\dot{p}$, the optimal value of $\dot{a}$ for $\Gamma'(\dot{p}, \dot{a})$ subject to $\dot{a} \le -\sigma_y|\dot{p}|_0$ is $\dot{a} = -\sigma_y|\dot{p}|_0$. Noting $I_S^*(\dot{p}, \dot{a}, \dot{p}) = \sigma_y|\dot{p}|_0$, we may thus restrict to the minimization of

$$\Gamma''(\dot{p}) = \sigma_y|\dot{p}|_0 - \sigma : \dot{p} - \mathbb{H}_{\text{iso}}\sigma_y a'|\dot{p}|_0 + \tau\mathbb{H}_{\text{iso}}\sigma_y^2|\dot{p}|_0^2 + \mathbb{H}_{\text{kin}} p' : \dot{p} + \tau\mathbb{H}_{\text{kin}}\dot{p} : \dot{p}.$$

For a minimizer, we have

$$0 \in -\sigma + 2\tau\mathbb{H}_{\rm iso}\sigma_y^2 \overset{\bullet}{p} + \mathbb{H}_{\rm kin} p' + 2\tau\mathbb{H}_{\rm kin}\overset{\bullet}{p} + \partial(\sigma_y - \mathbb{H}_{\rm iso}\sigma_y a')|\cdot|_0(\overset{\bullet}{p}).$$

Writing $\sigma = A - \tau\mathbb{C}\overset{\bullet}{p}$ we have

$$A - \mathbb{H}_{\rm kin} p' - \tau(\mathbb{C} + 2\mathbb{H}_{\rm iso}\sigma_y^2 + 2\tau\mathbb{H}_{\rm kin})\overset{\bullet}{p} \in \partial\sigma_y(1 - \mathbb{H}_{\rm iso}a')|\cdot|_0(\overset{\bullet}{p}).$$

Lemma 11.3 implies the asserted formula for $\overset{\bullet}{p}$. □

The formula of the proposition defines the stress function $\mathscr{S}^k : u^k \mapsto \sigma^k$ via

$$\widehat{\mathscr{S}}^k(u^k) = \begin{bmatrix} \mathscr{S}^k(u^k) \\ \mathscr{S}_1^k(u^k) \\ \mathscr{S}_2^k(u^k) \end{bmatrix} = \begin{bmatrix} \mathbb{C}\big(\varepsilon(u^k) - p^{k-1}\big) - \tau\mathbb{C}d_t p^k \\ -\mathbb{H}_{\rm iso}(a^{k-1} - \tau\sigma_y|d_t p^k|) \\ -\mathbb{H}_{\rm kin}(b^{k-1} + \tau d_t p^k) \end{bmatrix}.$$

For the special cases of perfect plasticity and linear isotropic and kinematic hardening, the formula for the stress $\sigma^k = \mathscr{S}^k(u^k)$ can be simplified. We recall the abbreviation $A^k = \mathbb{C}\big(\varepsilon(u^k) - p^{k-1}\big)$.

Examples 11.6 (i) For perfect plasticity we have

$$\mathscr{S}^k(u^k) = A^k - (|\operatorname{dev} A^k| - \sigma_y)_+ \frac{\operatorname{dev} A^k}{|\operatorname{dev} A^k|}.$$

(ii) For isotropic hardening we have

$$\mathscr{S}^k(u^k) = A^k - \frac{(|\operatorname{dev} A^k| - \sigma_y(1 - \mathbb{H}_{\rm iso}a^{k-1}))_+}{1 + \mathbb{H}_{\rm iso}\sigma_y^2/\mu} \frac{\operatorname{dev} A^k}{|\operatorname{dev} A^k|}.$$

(iii) For kinematic hardening we have

$$\mathscr{S}^k(u^k) = A^k - \frac{(|\operatorname{dev} A^k - \mathbb{H}_{\rm kin} p^{k-1}| - \sigma_y)_+}{1 + \mathbb{H}_{\rm kin}/\mu} \frac{\operatorname{dev} A^k - \mathbb{H}_{\rm kin} p^{k-1}}{|\operatorname{dev} A^k - \mathbb{H}_{\rm kin} p^{k-1}|}.$$

Remarks 11.7 (i) The operator $\mathscr{S}^k$ can be written as

$$\mathscr{S}^k(u^k) = A^k - 2\mu\tau d_t p^k$$

with

$$d_t p^k = (1/r)\big(|\operatorname{dev} A^k + M^{k-1}| - s^{k-1}\big)_+ \frac{\operatorname{dev} A^k + M^{k-1}}{|\operatorname{dev} A^k + M^{k-1}|}$$

for $M^k = -\mathbb{H}_{\rm kin} p^{k-1}$, $s^{k-1} = \sigma_y(1 - \mathbb{H}_{\rm iso}a^{k-1})$, and $r = 2\mu\tau(1 + \mathbb{H}_{\rm kin}/\mu + \mathbb{H}_{\rm iso}\sigma_y^2/\mu)$. Notice that M^{k-1} is deviatoric, i.e., $M^{k-1} = \operatorname{dev} M^{k-1}$.

(ii) If $\max\{\mathbb{H}_{\text{iso}}, \mathbb{H}_{\text{kin}}\} > 0$, then the mapping $\mathscr{S}^k$ is Lipschitz continuous and strongly monotone.

A time step of the discretized elastoplastic evolution problem can now be formulated in terms of displacement.

Corollary 11.2 (Displacement formulation) *The tuple*

$$(u^k, p^k, a^k, b^k) \in H^1(\Omega; \mathbb{R}^d) \times L^2(\Omega; \mathbb{R}^{d\times d}_{\text{sym}}) \times L^2(\Omega) \times L^2(\Omega; \mathbb{R}^{d\times d})$$

satisfies $u^k|_{\Gamma_{\text{D}}} = 0$*,*

$$\int_\Omega \mathbb{C}\big(\varepsilon(u^k) - p^k\big) : \varepsilon(v)\,\mathrm{d}x = \int_\Omega f(t_k)\cdot v\,\mathrm{d}x + \int_{\Gamma_{\text{N}}} g(t_k)\cdot v\,\mathrm{d}s$$

for all $v \in H^1(\Omega; \mathbb{R}^d)$ *with* $v|_{\Gamma_{\text{D}}} = 0$*, and*

$$\Sigma^k = \begin{bmatrix} \mathbb{C}\big(\varepsilon(u^k) - p^k\big) \\ \alpha^k \\ \beta^k \end{bmatrix} \in \partial I_S(d_t P^k) = \partial I_S\left(\begin{bmatrix} d_t p^k \\ d_t a^k \\ d_t b^k \end{bmatrix}\right)$$

if and only if

$$\int_\Omega \mathscr{S}^k(u^k) : \nabla v\,\mathrm{d}x = \int_\Omega f(t_k)\cdot v\,\mathrm{d}x + \int_{\Gamma_{\text{N}}} g(t_k)\cdot v\,\mathrm{d}x$$

for all $v \in H^1(\Omega; \mathbb{R}^d)$ *with* $v|_{\Gamma_{\text{D}}} = 0$ *and* (p^k, a^k, b^k) *are defined according to Proposition 11.7.*

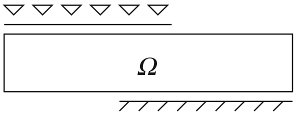
Remark 11.8 Notice that the stress function $\mathscr{S}^k$ only depends on ∇u^k, i.e., we may write $\mathscr{S}^k(u^k) = \mathscr{S}^k(\nabla u^k)$.

11.3.2 Newton Method for Nonlinear Elasticity

Given a stress function $\mathscr{S} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$, we consider the iterative solution and implementation of the problem of finding $u_h \in \mathscr{S}^1(\mathscr{T})^d$ such that $M(z)u_h(z) = u_{\text{D}}(z)$ for all $z \in \mathscr{N} \cap \Gamma_{\text{D}}$ and

$$\int_\Omega \mathscr{S}(\nabla u_h) : \nabla v_h\,\mathrm{d}x = \int_\Omega f\cdot v_h\,\mathrm{d}x + \int_{\Gamma_{\text{N}}} g\cdot v_h\,\mathrm{d}s$$

for all $v_h \in \mathscr{S}^1(\mathscr{T})^d$ with $M(z)v_h(z) = 0$ for all $z \in \mathscr{N} \cap \Gamma_{\text{D}}$. The matrix field $M : \Gamma_{\text{D}} \to \mathbb{R}^{d\times d}$ allows us to impose Dirichlet conditions on individual components of the vector field u_h on Γ_{D}, or to formulate gliding boundary conditions.



$$M(x,1) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \qquad u_{\mathrm{D}}(x,1) = \begin{bmatrix} 0 \\ \omega \end{bmatrix}$$

$$M(x,0) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad u_{\mathrm{D}}(x,0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Fig. 11.4 Partial Dirichlet boundary conditions and full Dirichlet boundary conditions specified by a matrix field M and a vector field u_{D}

Example 11.7 Consider $\Omega = (0,5)\times(0,1)$ and the Dirichlet conditions $u|_{[2,5]\times\{0\}} = 0$ and $u_2|_{[0,3]\times\{1\}} = \omega$ for a displacement field $u = (u_1, u_2)$. This is equivalent to requiring $Mu = u_{\mathrm{D}}$ on Γ_{D} with M and u_{D} specified in Fig. 11.4.

The application of the Newton scheme to the nonlinear system of equations leads to the following algorithm.

Algorithm 11.3 (*Newton scheme for nonlinear elasticity*) Let $u_h^0 \in \mathscr{S}^1(\mathscr{T})^d$ with $M(z)u_h^0(z) = u_{\mathrm{D}}(z)$ for all $z \in \mathscr{N} \cap \Gamma_{\mathrm{D}}$. Define the sequence $(u_h^\ell)_{\ell=0,1,\dots}$ by computing for $\ell = 1, 2, \dots$, a function $u_h^\ell \in \mathscr{S}^1(\mathscr{T})^d$ with $M(z)u_h^\ell(z) = u_{\mathrm{D}}(z)$ for all $z \in \mathscr{N} \cap \Gamma_{\mathrm{D}}$ and

$$\begin{aligned}\int_\Omega D\mathscr{S}(\nabla u_h^{\ell-1})[\nabla u_h^\ell] : \nabla v_h \,\mathrm{d}x = &\int_\Omega D\mathscr{S}(\nabla u_h^{\ell-1})[\nabla u_h^{\ell-1}] : \nabla v_h \,\mathrm{d}x \\ &- \int_\Omega \mathscr{S}(\nabla u_h^{\ell-1}) : \nabla v_h \,\mathrm{d}x \\ &+ \int_\Omega f \cdot v_h \,\mathrm{d}x + \int_{\Gamma_{\mathrm{N}}} g \cdot v_h \,\mathrm{d}s\end{aligned}$$

for all $v_h \in \mathscr{S}^1(\mathscr{T})^d$ with $M(z)v_h(z) = 0$ for all $z \in \mathscr{N} \cap \Gamma_{\mathrm{D}}$. Stop the iteration if $\|\nabla(u_h^\ell - u_h^{\ell-1})\|_{D\mathscr{S}(u_h^{\ell-1})} \le \varepsilon_{\mathrm{stop}}$ for $\|w_h\|^2_{D\mathscr{S}(u_h)} = \int_\Omega D\mathscr{S}(\nabla u_h)[\nabla w_h] : \nabla w_h \,\mathrm{d}x$.

We employ the basis $(\psi_{(z,p)} = e_p \varphi_z : z \in \mathscr{N}, p = 1, 2, \dots, d)$ of $\mathscr{S}^1(\mathscr{T})^d$ with the canonical basis vectors $e_p \in \mathbb{R}^d$, $p = 1, 2, \dots, d$. Given a vector field $u_h \in \mathscr{S}^1(\mathscr{T})^d$ with coefficient vector $U \in \mathbb{R}^{dL}$ with $L = \#\mathscr{N}$, we define the vector $F(U) \in \mathbb{R}^{dL}$

$$\big[F(U)\big]_{(z,p)} = \int_\Omega \mathscr{S}(\nabla u_h) : \nabla \psi_{(z,p)} \,\mathrm{d}x$$

for $z \in \mathscr{N}$ and $1 \le p \le d$. Similarly, we define the matrix $DF(U) \in \mathbb{R}^{dL\times dL}$ by

$$\big[DF(U)\big]_{(z,p),(y,q)} = \int_\Omega D\mathscr{S}(\nabla u_h)[\nabla \psi_{(z,p)}] : \nabla \psi_{(y,q)} \,\mathrm{d}x$$

for $z, y \in \mathcal{N}$ and $1 \leq p, q \leq d$. The full-rank matrix B is obtained by arranging the $d \times d$ matrices $M(z)$ for all $z \in \mathcal{N} \cap \Gamma_{\mathrm{D}}$ on the diagonal of a matrix $\widehat{B}$, and then deleting vanishing rows. Similarly, we arrange the vectors $u_{\mathrm{D}}(z) \in \mathbb{R}^d$, $z \in \mathcal{N} \cap \Gamma_{\mathrm{D}}$, in a vector $\widehat{W}$ and then delete the entries corresponding to deleted rows in $\widehat{B}$ to obtain a vector W. With these definitions, a step of the Newton iteration can be rewritten as

$$\begin{bmatrix} DF(U^{\ell-1}) & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} U^\ell \\ \Lambda \end{bmatrix} = \begin{bmatrix} DF(U^{\ell-1})U^{\ell-1} - F(U^{\ell-1}) \\ W \end{bmatrix}.$$

In our implementation a matrix $A \in \mathbb{R}^{d \times d}$ is identified with a vector $\widetilde{A} \in \mathbb{R}^{d^2}$ via $A_{ij} = \widetilde{A}_{(i-1)d+j}$, i.e.,

$$A = \begin{bmatrix} a_1 & \cdots & a_d \\ \vdots & \ddots & \vdots \\ a_{(d-1)d+1} & \cdots & a_{d^2} \end{bmatrix} \sim \quad \widetilde{A} = [a_1, a_2, \ldots, a_{d^2}]^\top.$$

The MATLAB codes displayed in Figs. 11.5 and 11.6 realize Algorithm 11.3. The code in Fig. 11.5 is an implementation of the Newton iteration and the routine displayed in Fig. 11.6 computes the required matrix $DF(u_h)$ and vector $F(u_h)$ for an iteration step in the Newton scheme. The routine also provides the stress field $\sigma_h = \mathcal{S}(\nabla u_h)$. In the implementation, the elementwise constant Jacobian matrix ∇u_h is contained in the array `Du` whose dimension is $\#\mathcal{T} \times d^2$. In a loop over all elements in $\mathcal{T}$, the elementwise contributions to the matrix $DF(u_h)$ and the vector $F(u_h)$ are computed, where the arrays `D_psi_1` and `D_psi_2` represent gradients of elements in the basis of $\mathcal{S}^1(\mathcal{T}_h)^d$. The routines `element_geometry.m` and `side_geometry.m` compute elementwise gradients of basis functions and volumes and midpoints of elements and sides, respectively.

11.3.3 Implementation of Elastoplasticity

The MATLAB routine for solving the quasi-stationary elastoplastic model problem displayed in Fig. 11.7 is based on the displacement formulation of Corollary 11.2 of a time step in the implicit discretization of the problem. We consider the k-th time step of the discretized elastoplastic evolution problem in the following. According to Remark 11.7, the nonlinear stress function $\mathcal{S}^k(u) = \mathcal{S}^k(\nabla u)$ is for a displacement field u and $A = \mathbb{C}(\varepsilon(u) - p^{k-1})$ given by

$$\mathcal{S}^k(\nabla u) = A - 2\mu\tau d_t p^k$$

with

$$d_t p^k = \frac{1}{2\mu\tau r}\big(|\operatorname{dev} A + M^{k-1}| - s^{k-1}\big)_+ \frac{\operatorname{dev} A + M^{k-1}}{|\operatorname{dev} A + M^{k-1}|}.$$

```
function nonlinear_elasticity(d_tmp,red)
global d; d = d_tmp; factor = 10;
[c4n,n4e,Db,Nb] = triang_beam(d,5);
for j = 1:red
    [c4n,n4e,Db,Nb] = red_refine(c4n,n4e,Db,Nb);
end
[nC,d] = size(c4n); u = zeros(d*nC,1);
[B,W] = diri_constraint(c4n,Db);
[DF,F,sigma] = nonlinear_fe_matrices(c4n,n4e,Nb,u);
corr = 1; eps_stop = 1e-4;
while corr > eps_stop
    b = [DF*u-F;W]; A = [DF,B';B,sparse(size(B,1),size(B,1))];
    x = A\b; u_new = x(1:d*nC);
    corr = sqrt((u-u_new)'*DF*(u-u_new)); u = u_new;
    [DF,F,sigma] = nonlinear_fe_matrices(c4n,n4e,Nb,u);
end
def = c4n+factor*reshape(u,d,nC)';
mod_sigma = sum(sigma.^2,2).^(1/2);
show_p1_def(c4n,n4e,def,mod_sigma);

function [B,W] = diri_constraint(c4n,Db)
[nC,d] = size(c4n); dNodes = unique(Db); nDb = size(dNodes,1);
B = sparse(d*nDb,d*nC); W = sparse(d*nDb,1);
for j = 1:nDb
    [M,U] = u_D(c4n(dNodes(j),:));
    B((j-1)*d+(1:d),(dNodes(j)-1)*d+(1:d)) = M;
    W((j-1)*d+(1:d)) = U;
end
essential_DNodes = find(sum(abs(B),2));
B = B(essential_DNodes,:); W = W(essential_DNodes,:);

function [M,U] = u_D(x)
d = size(x,2); M = zeros(d,d); U = zeros(d,1);
if x(d) == 0
    M = eye(d);
elseif x(d) == 1
    M(d,d) = 1; U(d) = -.025;
end
```

Fig. 11.5 Newton scheme for the solution of the nonlinear elasticity problem as in Algorithm 11.3

The derivative with respect to u is for v and $B = \mathbb{C}\varepsilon(v)$ given by

$$D\mathscr{S}^k(\nabla u)[\nabla v] = B$$

if $|\operatorname{dev} A + M^{k-1}| < s^{k-1}$ and

$$\begin{aligned} D\mathscr{S}^k(\nabla u)[\nabla v] = B &- \frac{1}{r}\big(|\operatorname{dev} A + M^{k-1}| - s^{k-1}\big)\frac{\operatorname{dev} B}{|\operatorname{dev} A + M^{k-1}|} \\ &+ \frac{s^{k-1}}{r}(\operatorname{dev} A + M^{k-1}) : \operatorname{dev} B \frac{\operatorname{dev} A + M^{k-1}}{|\operatorname{dev} A + M^{k-1}|^3} \end{aligned}$$

```
function [DF,F,sigma] = nonlinear_fe_matrices(c4n,n4e,Nb,u)
[nC,d] = size(c4n); nE = size(n4e,1); nNb = size(Nb,1);
F = zeros(d*nC,1); sigma = zeros(nE,d^2); Du = zeros(nE,d^2);
ctr_max = d^2*(d+1)^2*nE; ctr = 0;
I = zeros(ctr_max,1); J = zeros(ctr_max,1); X = zeros(ctr_max,1);
for k = 1:d
    Du(:,(k-1)*d+(1:d)) = comp_gradient(c4n,n4e,u(k:d:nC*d));
end
for j = 1:nE
    [mp_T,vol_T,grads_T] = geometry_element(c4n(n4e(j,:),:));
    Du_T = Du(j,:); sigma(j,:) = stress(Du_T,j);
    for m = 1:d+1
        for p = 1:d
            D_psi_A = zeros(1,d^2);
            D_psi_A(d*(p-1)+(1:d)) = grads_T(m,:);
            D_Sigma_T_A = stress_derivative(Du_T,D_psi_A,j);
            for n = 1:d+1
                for q = 1:d
                    ctr = ctr+1;
                    D_psi_B = zeros(1,d^2);
                    D_psi_B(d*(q-1)+(1:d)) = grads_T(n,:);
                    I(ctr) = d*n4e(j,m)-d+p;
                    J(ctr) = d*n4e(j,n)-d+q;
                    X(ctr) = vol_T*D_Sigma_T_A*D_psi_B';
            end; end
            phi_mp_T = zeros(d,1); phi_mp_T(p) = 1/(d+1);
            F(d*(n4e(j,m)-1)+p) = F(d*(n4e(j,m)-1)+p) ...
                +vol_T*(sigma(j,:)*D_psi_A'-f(mp_T)*phi_mp_T);
end; end; end
DF = sparse(I,J,X,d*nC,d*nC);
for j = 1:nNb
    [mp_S,vol_S] = geometry_side(c4n(Nb(j,:),:));
    for m = 1:d
        for p = 1:d
            phi_mp_S = zeros(d,1); phi_mp_S(p) = 1/d;
            F(d*(Nb(j,m)-1)+p) = F(d*(Nb(j,m)-1)+p)...
                -vol_S*g(mp_S)*phi_mp_S;
end; end; end

function val = f(x); global d; val = zeros(1,d);
function val = g(x); global d; val = zeros(1,d);

function val = stress(Du_T,j)
global d; transp = reshape(1:d^2,d,d)';
E_sym = (Du_T+Du_T(transp(:)))/2;
val = .1*sum(E_sym.^2,2)*E_sym+E_sym;

function val = stress_derivative(Du_T,B,j)
global d; transp = reshape(1:d^2,d,d)';
E_sym = (Du_T+Du_T(transp(:)))/2; B_sym = (B+B(transp(:)))/2;
val = .2*sum(E_sym.*B_sym,2)*E_sym+.1*sum(E_sym.^2,2)*B_sym+B_sym;
```

Fig. 11.6 Finite element matrices and vectors required in the Newton iteration of Algorithm 11.3 with stress function defined through the energy density $W(E) = |E|^4/40 + |E|^2/2$

```
function elastoplasticity(d_tmp,red)
global d p_old p_new a_old a_new tau t; d = d_tmp;
[c4n,n4e,Db,Nb] = triang_beam(d,5);
shift_vec = [2.5,.5,.5];
c4n = 0.01*(c4n-ones(size(c4n,1),1)*shift_vec(1:d));
for j = 1:red
    [c4n,n4e,Db,Nb] = red_refine(c4n,n4e,Db,Nb);
end
tau = 2^(-red)/40;
[nC,d] = size(c4n); nE = size(n4e,1);
T = .2; factor = 10; material_parameters();
p_old = zeros(nE,d^2); p_new = zeros(nE,d^2);
a_old = zeros(nE,1); a_new = zeros(nE,1);
for k = 1:floor(T/tau)
    t = k*tau;
    [u,sigma] = plastic_step(c4n,n4e,Db,Nb);
    p_old = p_new; a_old = a_new;
    def = c4n+factor*reshape(u,d,nC)';
    mod_sigma = sum(sigma.^2,2).^(1/2);
    mod_p = sum(p_new.^2,2).^(1/2);
    figure(1); show_p1_def(c4n,n4e,def,mod_sigma);
    figure(2); show_p1_def(c4n,n4e,def,mod_p);
end

function [u,sigma] = plastic_step(c4n,n4e,Db,Nb)
[nC,d] = size(c4n); u = zeros(d*nC,1);
[B,W] = diri_constraint(c4n,Db);
[DF,F,sigma] = nonlinear_fe_matrices(c4n,n4e,Nb,u);
corr = 1; eps_stop = 1e-2;
while corr > eps_stop
    b = [DF*u-F;W];
    A = [DF,B';B,sparse(size(B,1),size(B,1))];
    x = A\b;
    u_new = x(1:d*nC);
    corr = sqrt((u-u_new)'*DF*(u-u_new))
    u = u_new;
    [DF,F,sigma] = nonlinear_fe_matrices_plast(c4n,n4e,Nb,u);
end

function material_parameters()
global nu lambda mu sigma_y H_kin H_iso;
E = 1.37e+11; nu = 0.3;
lambda = nu*E/((1+nu)*(1-2*nu)); mu = E/(2*(1+nu));
sigma_y = 4.5e08; H_kin = 1/1000; H_iso = 0;
```

Fig. 11.7 The implicit discretization of the elastoplastic evolution problem leads to a nonlinearly elastic problem in every time step that is solved with a Newton iteration; the defined material parameters model realistic steel

otherwise. We use global variables to avoid long arguments of function calls. This enables us to use the routine that assembles the matrices for the Newton scheme with the stress function $\mathscr{S}^k$. The routine `nonlinear_fe_matrices_plast.m` is thus a copy of the routine `nonlinear_fe_matrices.m` in which the subroutines,

```
function val = stress(Du_T,j)
global d lambda mu sigma_y H_kin H_iso;
global tau p_old p_new a_old a_new;
transp = reshape(1:d^2,d,d)';
id_mat = reshape(eye(d),1,d^2);
E_sym = (Du_T+Du_T(transp(:)))/2;
A = lambda*tr(E_sym-p_old(j,:))*id_mat+2*mu*(E_sym-p_old(j,:));
M = -H_kin*p_old(j,:);
r = 1+H_iso*sigma_y^2/mu+H_kin/mu;
s = sigma_y*(1-H_iso*a_old(j));
dt_p = zeros(1,d^2);
if norm(dev(A+M))-s > 0
    dt_p = (1/(2*mu*tau*r))*(1-s/norm(dev(A+M)))*dev(A+M);
end
p_new(j,:) = p_old(j,:)+tau*dt_p;
a_new(j) = a_old(j)-tau*sigma_y*norm(dt_p);
val = A-2*mu*tau*dt_p;

function dev_A = dev(A)
global d; dev_A = A-tr(A)/d*reshape(eye(d),1,d^2);
function tr_A = tr(A)
global d; tr_A = sum(A(1:d+1:d^2),2);
```

```
function val = stress_derivative(Du_T,Dv,j)
global d lambda mu sigma_y H_kin H_iso;
global p_old a_old;
transp = reshape(1:d^2,d,d)';
id_mat = reshape(eye(d),1,d^2);
E_sym = (Du_T+Du_T(transp(:)))/2;
A = lambda*tr(E_sym-p_old(j,:))*id_mat+2*mu*(E_sym-p_old(j,:));
M = -H_kin*p_old(j,:);
r = 1+H_iso*sigma_y^2/mu+H_kin/mu;
s = sigma_y*(1-H_iso*a_old(j));
Dv_sym = (Dv+Dv(transp(:)))/2;
B = lambda*tr(Dv_sym)*id_mat+2*mu*Dv_sym;
val = B;
if norm(dev(A+M))-s > 0
    val = B+(1/r)*(norm(dev(A+M))-s)*dev(B)/norm(dev(A+M))...
        -(1/r)*s*(dev(A+M)*dev(B)')*dev(A+M)/norm(dev(A+M))^3;
end

function dev_A = dev(A)
global d; dev_A = A-tr(A)/d*reshape(eye(d),1,d^2);
function tr_A = tr(A)
global d; tr_A = sum(A(1:d+1:d^2),2);
```

Fig. 11.8 Implementation of the stress function $\mathscr{S}^k(\nabla u)$ for general von Mises plasticity and its derivative $D\mathscr{S}^k(\nabla u)[\nabla v]$; the index j refers to the corresponding element in the triangulation

defining the stress and its derivative, have been eliminated. These are replaced by the functions displayed in Fig. 11.8. The argument j in the calls of the stress function and its derivative in the assembly of the matrices allows us to access the elementwise values of globally defined fields in the subroutines. We remark that the variables β

and b are eliminated from the problem in the implementation via the identities $p = b$ and $\beta = -\mathbb{H}_{kin} b$.

References

1. Alberty, J., Carstensen, C., Zarrabi, D.: Adaptive numerical analysis in primal elastoplasticity with hardening. Comput. Methods Appl. Mech. Eng. **171**(3–4), 175–204 (1999). http://dx.doi.org/10.1016/S0045-7825(98)00210-2
2. Bartels, S.: Quasi-optimal error estimates for implicit discretizations of rate-independent evolutions. SIAM J. Numer. Anal. **52**(2), 708–716 (2014). http://dx.doi.org/10.1137/130933964
3. Bartels, S., Mielke, A., Roubíček, T.: Quasi-static small-strain plasticity in the limit of vanishing hardening and its numerical approximation. SIAM J. Numer. Anal. **50**(2), 951–976 (2012). http://dx.doi.org/10.1137/100819205
4. Dal Maso, G., DeSimone, A., Mora, M.G.: Quasistatic evolution problems for linearly elastic-perfectly plastic materials. Arch. Ration. Mech. Anal. **180**(2), 237–291 (2006). http://dx.doi.org/10.1007/s00205-005-0407-0
5. Han, W., Reddy, B.D.: Plasticity. Interdisciplinary Applied Mathematics, 2nd edn. Springer, New York (2013) Mathematical Theory and Numerical Analysis
6. Johnson, C.: On plasticity with hardening. J. Math. Anal. Appl. **62**(2), 325–336 (1978)
7. Mielke, A.: Evolution of Rate-Independent Systems. In: Evolutionary Equations. Vol. II, Handbook of Differential Equations, pp. 461–559. Elsevier/North-Holland, Amsterdam (2005)
8. Sauter, M., Wieners, C.: On the superlinear convergence in computational elasto-plasticity. Comput. Methods Appl. Mech. Eng. **200**(49–52), 3646–3658 (2011). http://dx.doi.org/10.1016/j.cma.2011.08.011
9. Simo, J.C., Hughes, T.J.R.: Computational Inelasticity. Interdisciplinary Applied Mathematics, vol. 7. Springer, New York (1998)
10. Stefanelli, U.: A variational principle for hardening elastoplasticity. SIAM J. Math. Anal. 40(2), 623–652 (2008). http://dx.doi.org/10.1137/070692571

Appendix A
Auxiliary Routines

A.1 Triangulations

A.1.1 Domains in $\mathbb{R}^d$

Triangulations of some domains in $\mathbb{R}^d$ are provided by the routines displayed in Figs. A.1 and A.2. The routine `triang_cube.m` defines a coarse triangulation of the d-dimensional unit cube for $d = 1, 2, 3$ with a partition of the boundary into Dirichlet and Neumann parts specified by

$$\Omega = (0,1)^d, \quad \Gamma_{\mathrm{D}} = \partial\Omega \cap \big(\mathbb{R}^{d-1} \times \{0\}\big), \quad \Gamma_{\mathrm{N}} = \partial\Omega \setminus \Gamma_{\mathrm{D}}.$$

A uniform triangulation of a two-dimensional strip with side lengths $L \in \mathbb{N}$ and 1 into $2L$ right isosceles triangles and the Dirichlet part of the boundary consisting of the ends of the strip, i.e.,

$$\Omega = (0, L) \times (0, 1), \quad \Gamma_{\mathrm{D}} = \{0, L\} \times [0, 1], \quad \Gamma_{\mathrm{N}} = \partial\Omega \setminus \Gamma_{\mathrm{D}},$$

is computed in the routine `triang_strip.m`. The routine `triang_beam.m` defines a uniform partition of the two- or three-dimensional beam

$$\Omega = (0, L) \times (0, 1)^{d-1}$$

with variable integer length $L > 0$. The boundary is partitioned according to

$$\Gamma_{\mathrm{D}} = \big\{(x_1, \ldots, x_d) \in \partial\Omega : (x_1, x_d) \in [0, 3] \times \{1\} \text{ or } (x_1, x_d) \in [2, L] \times \{0\}\big\}$$

and $\Gamma_{\mathrm{N}} = \partial\Omega \setminus \Gamma_{\mathrm{D}}$. Figure A.2 shows the MATLAB code `triang_ring.m` that provides an approximate triangulation of the annulus $\Omega = B_2(0) \setminus \overline{B_1(0)}$ with Dirichlet boundary $\Gamma_{\mathrm{D}} = \partial\Omega$ and empty Neumann boundary $\Gamma_{\mathrm{N}} = \emptyset$. The triangulation is obtained from a triangulation of the unit square via the parametrization

S. Bartels, *Numerical Methods for Nonlinear Partial Differential Equations*,
Springer Series in Computational Mathematics 47,
DOI 10.1007/978-3-319-13797-1

```
function [c4n,n4e,Db,Nb] = triang_cube(d)
if d == 1
    c4n = [0;1]; n4e = [1,2]; Db = 1; Nb = 2;
elseif d == 2
    c4n = [0,0;1,0;1,1;0,1]; n4e = [1,2,3;1,3,4];
    Db = [1,2]; Nb = [2,3;3,4;4,1];
elseif d == 3
    c4n = [0,0,0;1,0,0;1,1,0;0,1,0;0,0,1;1,0,1;1,1,1;0,1,1];
    n4e = [1,2,3,7;1,6,2,7;1,5,6,7;1,8,5,7;1,4,8,7;1,3,4,7];
    Db = [1,2,3;1,4,3];
    Nb = [2,3,7;2,7,6;1,2,6;1,6,5;5,6,7;1,8,5;5,7,8;1,4,8;...
        4,7,8;3,4,7];
end
```

```
function [c4n,n4e,Db,Nb] = triang_strip(L)
n4e = [[1:L;L+1+(2:L+1);L+1+(1:L)]';[1:L;2:L+1;L+1+(2:L+1)]'];
c4n = [[0:L;zeros(1,L+1)]';[0:L;ones(1,L+1)]'];
Db = [L+1,2*L+2;L+2,1];
Nb = [[1:L;2:L+1]';[2*L+2:-1:L+3;2*L+1:-1:L+2]'];
```

```
function [c4n,n4e,Db,Nb] = triang_beam(d,L)
if d == 2
    c4n_ref = [0,0;1,0;1,1;0,1]; n4e_ref = [1,2,3;1,3,4];
    bdy_left = [4,1]; bdy_mid = [1,2;3,4]; bdy_right = [2,3];
elseif d == 3
    c4n_ref = [0,0,0;0,0,1;0,1,0;0,1,1;1,0,0;1,0,1;1,1,0;1,1,1];
    n4e_ref = [1,2,8,4;1,2,6,8;1,3,4,8;1,3,8,7;1,5,7,8;1,5,8,6];
    bdy_left = [1,2,4;1,3,4]; bdy_right = [6,5,8;7,5,8];
    bdy_mid = [1,2,6;1,3,7;1,5,6;1,5,7;4,2,8;6,2,8;4,3,8;7,3,8];
end
nC = size(c4n_ref,1); c4n = c4n_ref; n4e = n4e_ref;
bdy = [bdy_left;bdy_mid];
for k = 2:L
    c4n = [c4n;[c4n_ref(:,1)+(k-1),c4n_ref(:,2:d)]];
    n4e = [n4e;n4e_ref+nC*(k-1)];
    bdy = [bdy;bdy_mid+nC*(k-1)];
end
bdy = [bdy;bdy_right+nC*(L-1)];
[c4n,¬,K] = unique(c4n,'rows','first');
n4e = K(n4e); bdy = K(bdy); Db = []; Nb = [];
for j = 1:size(bdy,1);
    mp_S = sum(c4n(bdy(j,:),:),1)/d;
    if (mp_S(1)≤3 && mp_S(d)==1) || (mp_S(1)≥2 && mp_S(d)==0)
        Db = [Db;bdy(j,:)];
    else
        Nb = [Nb;bdy(j,:)];
    end
end
```

Fig. A.1 Generation of triangulations of the cube $\Omega = (0, 1)^d$ (*top*), the strip $\Omega = (0, L) \times (0, 1)$ for $L \in \mathbb{N}$ (*middle*), and the beam $\Omega = (0, L) \times (0, 1)^{d-1}$ for $L \in \mathbb{N}$ (*bottom*)

```
function [c4n,n4e,Db,Nb] = triang_ring(red)
c4n_ref = [0,0;1,0;1,1;0,1]; n4e = [1,2,3;1,3,4];
Db = [2,3;4,1]; Nb = [1,2;3,4];
for j = 1:red
    [c4n_ref,n4e,Db,Nb] = red_refine(c4n_ref,n4e,Db,Nb);
end
idx = find(c4n_ref(:,2)==1); c4n_ref(idx,2) = 0;
[c4n_ref,¬,K] = unique(c4n_ref,'rows','first');
n4e = K(n4e); Db = K(Db); Nb = [];
r = c4n_ref(:,1); phi = c4n_ref(:,2);
c4n = [(r+1).*cos(2*pi*phi),(r+1).*sin(2*pi*phi)];
```

Fig. A.2 Generation of an approximate triangulation of the annulus $B_2(0) \setminus \overline{B_1(0)}$

$$f : (0,1) \times [0,1] \to B_2(0) \setminus \overline{B_1(0)}, \quad (r,\phi) \mapsto (r+1)\big(\cos(2\pi\phi), \sin(2\pi\phi)\big).$$

Multiply occurring nodes in the image of the triangulation are eliminated with the help of the MATLAB command `unique`.

A.1.2 Hypersurfaces in $\mathbb{R}^3$

Discrete surfaces, i.e., unions of flat triangles in $\mathbb{R}^3$, that define Lipschitz continuous submanifolds in $\mathbb{R}^3$ are computed in the MATLAB routines displayed in Fig. A.3. Starting with a triangulation of the boundary of the cube $[-1,1]^3/\sqrt{3}$, an approximation of the unit sphere is obtained by alternatingly projecting the nodes onto the unit sphere and refining the triangulation. This is realized in the program `triang_sphere.m`. An approximation of the two-dimensional torus $T_{r,R}$ with radii $0 < r < R$ is computed in the routine `triang_torus.m` which employs the transformation $f : [0,2\pi]^2 \to \mathbb{R}^3$ defined by

$$f(u,v) = [(R + r\cos(v))\cos(u), (R + r\cos(v))\sin(u), r\sin(v)]^\top.$$

The surface is closed and the boundary parts Γ_D and Γ_N are empty.

```
function [n4e,c4n,Db,Nb] = triang_sphere(red)
c4n = [-1,-1,-1;1,-1,-1;1,1,-1;-1,1,-1;-1,-1,1;1,-1,1;...
    1,1,1;-1,1,1]/sqrt(3);
n4e = [1,2,6;6,5,1;2,3,7;7,6,2;3,8,7;3,4,8;4,5,8;4,1,5;...
    6,7,8;6,8,5;1,3,2;1,4,3];
Db = []; Nb = [];
for j = 1:red
    [c4n,n4e,Db,Nb] = red_refine_surf(c4n,n4e,Db,Nb);
    c4n = c4n./(sqrt(sum(c4n.^2,2))*[1,1,1]);
end
```

```
function [c4n,n4e,Db,Nb] = triang_torus(r,R,red)
c4n_ref = [0,0;1,0;1,1;0,1]; n4e = [1,2,3;1,3,4];
Db = []; Nb = [];
for j = 1:red
    [c4n_ref,n4e,Db,Nb] = red_refine(c4n_ref,n4e,Db,Nb);
end
idx = find(c4n_ref(:,2)==1); c4n_ref(idx,2) = 0;
idx = find(c4n_ref(:,1)==1); c4n_ref(idx,1) = 0;
[c4n_ref,¬,K] = unique(c4n_ref,'rows','first');
n4e = K(n4e);
u = 2*pi*c4n_ref(:,1); v = 2*pi*c4n_ref(:,2);
c4n = [(R+r*cos(v)).*cos(u),(R+r*cos(v)).*sin(u),r*sin(v)];
```

Fig. A.3 Discrete surfaces defined by approximate triangulations of the unit sphere (*top*) and the torus with radii $0 < r < R$ (*bottom*)

A.2 Grid Refinement

Coarse triangulations can be refined with the MATLAB routine `red_refine.m` displayed in Figs. A.4 and A.5. The refinement procedure partitions every d-dimensional simplex into 2^d subsimplices by bisecting its one-dimensional subsimplices and appropriately connecting new nodes as illustrated in Figs. A.6 and A.7. The routine also provides matrices that allow for computing the coefficients of a given finite element function on the coarse triangulation with respect to the nodal basis on the refined triangulation by a matrix vector multiplication. For continuous, piecewise affine functions this is realized with the matrix `P1` and for elementwise constant functions with the matrix `P0`. For triangulations of hypersurfaces in $\mathbb{R}^3$, the same strategy can be used to refine a given simplicial approximation of a surface, cf. Fig. A.8. The code `red_refine_surf.m` shown in Fig. A.9 is a straightforward modification of the routine `red_refine.m` that incorporates the additional coordinates of the nodes. A postprocessing procedure that projects the newly created nodes onto a given surface can be incorporated to increase the accuracy of the approximation.

```
function [c4nNew,n4eNew,DbNew,NbNew,P0,P1] ...
    = red_refine(c4n,n4e,Db,Nb)
[nC,d] = size(c4n); nE = size(n4e,1);
nDb = size(Db,1); nNb = size(Nb,1);
K = sparse(1:nC,1:nC,1:nC);
c4nNew = c4n; n4eNew = zeros(nE*2^d,d+1);
DbNew = zeros(nDb*2^(d-1),d); NbNew = zeros(nNb*2^(d-1),d);
P0 = sparse(2^d*nE,nE);
nr_nodes = nC;
for j = 1:nE
    for k = 1:d+1
        for m = 1:d+1
            if K(n4e(j,k),n4e(j,m))==0
                nr_nodes = nr_nodes+1;
                K(n4e(j,k),n4e(j,m)) = nr_nodes;
                K(n4e(j,m),n4e(j,k)) = nr_nodes;
                I1(2*(nr_nodes-nC-1)+(1:2)) = [nr_nodes,nr_nodes];
                I2(2*(nr_nodes-nC-1)+(1:2)) = [n4e(j,k),n4e(j,m)];
                EE(2*(nr_nodes-nC-1)+(1:2)) = [1 1]/2;
                c4nNew(nr_nodes,:) = ...
                    (c4n(n4e(j,k),:)+c4n(n4e(j,m),:))/2;
            end
        end
    end
    nodes = K(n4e(j,:),n4e(j,:));
    if d == 1
        n4eNew(2*(j-1)+(1:2),:) = ...
            [nodes(1,1),nodes(1,2);nodes(1,2),nodes(2,2)];
    elseif d == 2
        n4eNew(4*(j-1)+(1:4),:) = ...
            [nodes(1,1),nodes(1,2),nodes(1,3);...
            nodes(1,2),nodes(2,3),nodes(1,3);...
            nodes(1,2),nodes(2,2),nodes(2,3);...
            nodes(1,3),nodes(2,3),nodes(3,3)];
    elseif d == 3
        n4eNew(8*(j-1)+(1:8),:) = ...
            [nodes(1,1),nodes(1,2),nodes(1,3),nodes(1,4);...
            nodes(1,2),nodes(2,2),nodes(2,3),nodes(2,4);...
            nodes(1,3),nodes(2,3),nodes(3,3),nodes(3,4);...
            nodes(1,4),nodes(2,4),nodes(3,4),nodes(4,4);...
            nodes(1,2),nodes(1,3),nodes(1,4),nodes(2,4);...
            nodes(2,3),nodes(1,3),nodes(1,2),nodes(2,4);...
            nodes(1,3),nodes(1,4),nodes(2,4),nodes(3,4);...
            nodes(1,3),nodes(2,4),nodes(2,3),nodes(3,4)];
    end
    P0(2^d*(j-1)+(1:2^d),j) = ones(1,2^d);
end
I1 = [I1,1:nC]; I2 = [I2,1:nC]; EE = [EE,ones(1,nC)];
P1 = sparse(I1,I2,EE,nr_nodes,nC);
```

Fig. A.4 MATLAB implementation of a uniform refinement procedure that partitions every simplex into 2^d subsimplices (continued in Fig. A.5)

```
for j = 1:nDb
    nodes = K(Db(j,:),Db(j,:));
    if d == 1
        DbNew = Db;
    elseif d == 2
        DbNew(2*(j-1)+(1:2),:) = ...
            [nodes(1,1),nodes(1,2);...
            nodes(1,2),nodes(2,2)];
    elseif d == 3
        DbNew(4*(j-1)+(1:4),:) = ...
            [nodes(1,1),nodes(1,2),nodes(1,3);...
            nodes(1,2),nodes(2,2),nodes(2,3);...
            nodes(1,2),nodes(2,3),nodes(1,3);...
            nodes(1,3),nodes(2,3),nodes(3,3)];
    end
end
for j = 1:nNb
    nodes = K(Nb(j,:),Nb(j,:));
    if d == 1
        NbNew = Nb;
    elseif d == 2
        NbNew(2*(j-1)+(1:2),:) = ...
            [nodes(1,1),nodes(1,2);...
            nodes(1,2),nodes(2,2)];
    elseif d == 3
        NbNew(4*(j-1)+(1:4),:) = ...
            [nodes(1,1),nodes(1,2),nodes(1,3);...
            nodes(1,2),nodes(2,2),nodes(2,3);...
            nodes(1,2),nodes(2,3),nodes(1,3);...
            nodes(1,3),nodes(2,3),nodes(3,3)];
    end
end
```

Fig. A.5 MATLAB implementation of a uniform refinement procedure that partitions every simplex into 2^d subsimplices (continued from Fig. A.4)

Fig. A.6 Partitioning of one- and two-dimensional simplices to define refined triangulations

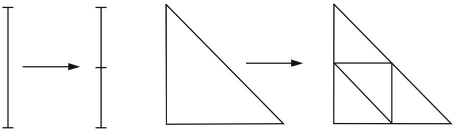
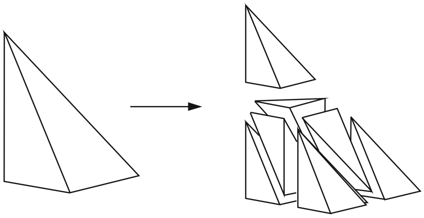


Fig. A.7 Partitioning of the three-dimensional simplex to define refined triangulations



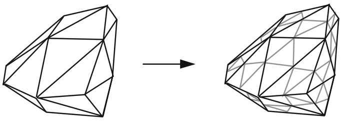


Fig. A.8 Uniform refinement of a triangulation of a discrete surface

```
function [c4nNew,n4eNew,DbNew,NbNew,P0,P1] ...
    = red_refine_surf(c4n,n4e,Db,Nb)
nC = size(c4n,1); nE = size(n4e,1);
nDb = size(Db,1); nNb = size(Nb,1);
K = sparse(1:nC,1:nC,1:nC);
c4nNew = c4n; n4eNew = zeros(nE*4,3);
DbNew = zeros(nDb*2,2); NbNew = zeros(nNb*2,2);
P0 = sparse(4*nE,nE); nr_nodes = nC;
for j = 1:nE
    for k = 1:3
        for m = 1:3
            if K(n4e(j,k),n4e(j,m))==0
                nr_nodes = nr_nodes+1;
                K(n4e(j,k),n4e(j,m)) = nr_nodes;
                K(n4e(j,m),n4e(j,k)) = nr_nodes;
                I1(2*(nr_nodes-nC-1)+(1:2)) = [nr_nodes,nr_nodes];
                I2(2*(nr_nodes-nC-1)+(1:2)) = [n4e(j,k),n4e(j,m)];
                EE(2*(nr_nodes-nC-1)+(1:2)) = [1 1]/2;
                c4nNew(nr_nodes,:) = ...
                    (c4n(n4e(j,k),:)+c4n(n4e(j,m),:))/2;
            end
        end
    end
    nodes = K(n4e(j,:),n4e(j,:));
    n4eNew(4*(j-1)+(1:4),:) = ...
        [nodes(1,1),nodes(1,2),nodes(1,3);...
        nodes(1,2),nodes(2,3),nodes(1,3);...
        nodes(1,2),nodes(2,2),nodes(2,3);...
        nodes(1,3),nodes(2,3),nodes(3,3)];
    P0(4*(j-1)+(1:4),j) = ones(4,1);
end
I1 = [I1,1:nC]; I2 = [I2,1:nC]; EE = [EE,ones(1,nC)];
P1 = sparse(I1,I2,EE,nr_nodes,nC);
for j = 1:nDb
    nodes = K(Db(j,:),Db(j,:));
    DbNew(2*(j-1)+(1:2),:) = ...
        [nodes(1,1),nodes(1,2);nodes(1,2),nodes(2,2)];
end
for j = 1:nNb
    nodes = K(Nb(j,:),Nb(j,:));
    NbNew(2*(j-1)+(1:2),:) = ...
        [nodes(1,1),nodes(1,2);nodes(1,2),nodes(2,2)];
end
```

Fig. A.9 MATLAB implementation of a uniform refinement procedure for simplicial surfaces in $\mathbb{R}^3$; *every flat triangle* in $\mathbb{R}^3$ is partitioned into 4 *subtriangles*

A.3 Visualization

A.3.1 Displaying Commands

Table A.1 lists some MATLAB commands which plot functions and manipulate a figure. Detailed explanations are available from MATLAB's help function.

A.3.2 Functions and Vector Fields

The routines shown in Fig. A.10 visualize functions and vector fields that are continuous and piecewise affine. The first routine `show_p1.m` displays the domain $\Omega \subset \mathbb{R}^d$ that is colored by the values of the scalar finite element function $u_h : \Omega \to \mathbb{R}$. The command `drawnow` enforces an immediate update of graphical output which is required when evolution problems are solved numerically. Three-dimensional vector fields $u_h : \Omega \to \mathbb{R}^3$ on domains $\Omega \subset \mathbb{R}^d$ with $d = 1, 2, 3$ are visualized with the routine `show_p1_field.m`. A deformed domain defined by $u_h(\Omega)$ for $\Omega \subset \mathbb{R}^d$ and $u_h : \Omega \to \mathbb{R}^d$ is displayed with the program `show_p1_def.m`. An optional elementwise constant quantity defines a coloring of the deformed domain. Parametrized surfaces in $\mathbb{R}^3$ defined by a mapping $u_h : \Omega \to \mathbb{R}^3$ are visualized with the routine `show_p1_para.m`. Discrete surfaces defined by unions of flat triangles together with continuous, elementwise affine functions on the surface are plotted with the MATLAB code `show_p1_surf.m` displayed in Figs. A.11.

Table A.1 MATLAB commands that generate and manipulate plots and figures

`plot, plot3`	Plots a polygonal curve in $\mathbb{R}^2$ or $\mathbb{R}^3$
`trimesh`	Displays a triangulation in $\mathbb{R}^2$
`tetramesh`	Displays a triangulation in $\mathbb{R}^3$
`trisurf`	Shows the graph of a scalar function on a triangulation (d=2)
`quiver, quiver3`	Plots a two- or three-dimensional vector field
`clf`	Clears a figure
`drawnow`	Updates a figure
`axis`	Sets the axes in a figure including the color range (optional)
`axis square`	Equal scaling of axes
`axis on/off`	Switches coordinate axes on or off
`colorbar`	Displays a color bar
`subplot`	Shows several plots in one figure
`view`	Changes the perspective
`colormap`	Chooses a color scale

```
function show_p1(c4n,n4e,Db,Nb,u)
d = size(c4n,2);
if d == 1
    plot(c4n(n4e),u(n4e));
elseif d == 2
    trisurf(n4e,c4n(:,1),c4n(:,2),u);
elseif d == 3
    trisurf([Db;Nb],c4n(:,1),c4n(:,2),c4n(:,3),u);
end
drawnow;
```

```
function show_p1_field(c4n,u)
[nC,d] = size(c4n);
X = [c4n,zeros(nC,3-d)];
quiver3(X(:,1),X(:,2),X(:,3),u(1:3:3*nC),u(2:3:3*nC),u(3:3:3*nC));
drawnow;
```

```
function show_p1_def(c4n,n4e,def,mod_sigma)
[nC,d] = size(def);
if d == 1
    plot(c4n(n4e),def(n4e));
elseif d == 2
    trisurf(n4e,def(:,1),def(:,2),zeros(nC,1),mod_sigma);
    view(0,90)
elseif d == 3
    tetramesh(n4e,def,mod_sigma);
end
drawnow;
```

```
function show_p1_para(c4n,n4e,u)
nC = size(c4n,1);
trisurf(n4e,u(0*nC+(1:3:3*nC)),u(3*nC+(1:3:3*nC)),...
    u(6*nC+(1:3:3*nC)));
drawnow;
```

Fig. A.10 MATLAB routines that visualize a scalar quantity $u_h : \Omega \to \mathbb{R}$ (*first*), a vector field $u_h : \Omega \to \mathbb{R}^3$ (*second*), a deformation $u_h : \Omega \to \mathbb{R}^d$ (*third*), or a parametrized surface $u_h : \Omega \to \mathbb{R}^3$ (*fourth routine*)

```
function show_p1_surf(c4n,n4e,H)
trisurf(n4e,c4n(:,1),c4n(:,2),c4n(:,3),H);
drawnow;
```

Fig. A.11 MATLAB routine that displays a discrete surface colored by a scalar quantity

A.4 Various Routines

A.4.1 Finite Element Gradient

The function `comp_gradient.m` shown in Fig. A.12 computes the elementwise constant gradient of a given continuous, elementwise affine function represented in

```
function du = comp_gradient(c4n,n4e,u)
d = size(c4n,2); nE = size(n4e,1);
du = zeros(nE,d);
for j = 1:nE
    X_T = [ones(1,d+1);c4n(n4e(j,:),:)'];
    grads_T = X_T\[zeros(1,d);eye(d)];
    du(j,:) = u(n4e(j,:))'*grads_T;
end
```

Fig. A.12 Computation of the elementwise vectorial values of the gradient of a $P1$ function

terms of its nodal values, i.e., the routine computes for a given function $u_h \in \mathscr{S}^1(\mathscr{T}_h)$ the matrix

$$\left[\nabla u_h|_{T_1}, \nabla u_h|_{T_2}, \dots, \nabla u_h|_{T_M}\right]^\top \in \mathbb{R}^{\#\mathscr{T}_h \times d}.$$

A.4.2 Element and Side Geometry

Given the local coordinates

$$Z_T = [z_1, z_2, \dots, z_{d+1}]^\top \in \mathbb{R}^{(d+1)\times d}, \quad Z_S = [z_1, z_2, \dots, z_d]^\top \in \mathbb{R}^{d\times d}$$

of an element or a side the MATLAB routines `geometry_element.m` and `geometry_side.m` displayed in Fig. A.13 compute the corresponding midpoint

```
function [mp_T,vol_T,grads_T] = geometry_element(Z_T)
d = size(Z_T,2);
mp_T = sum(Z_T,1)/(d+1);
X_T = [ones(1,d+1);Z_T'];
vol_T = det(X_T)/factorial(d);
grads_T = X_T\[zeros(1,d);eye(d)];
```

```
function [mp_S,vol_S] = geometry_side(Z_S)
d = size(Z_S,2);
mp_S = sum(Z_S,1)/d;
if d == 1
    vol_S = 1;
elseif d == 2
    vol_S = norm(Z_S(1,:)-Z_S(2,:));
elseif d == 3
    vol_S = norm(cross(Z_S(3,:)-Z_S(1,:),Z_S(2,:)-Z_S(1,:)),2)/2;
end
```

Fig. A.13 Determination of the midpoint and volume (*surface area*) of elements and sides; for elements (*top*) the gradients of the nodal basis functions associated to an element are provided

```
function A_v = average_quant_surf(c4n,n4e,v)
nC = size(c4n,1); nE = size(n4e,1);
area_patch = zeros(nC,1); quant_patch = zeros(nC,1);
for j = 1:nE
    n_T = cross(c4n(n4e(j,2),:)-c4n(n4e(j,1),:),...
        c4n(n4e(j,3),:)-c4n(n4e(j,2),:));
    area_T = norm(n_T)/2;
    area_patch(n4e(j,:)) = area_patch(n4e(j,:))+area_T;
    quant_patch(n4e(j,:)) = quant_patch(n4e(j,:))+area_T*v(j);
end
A_v = quant_patch./area_patch;
```

Fig. A.14 Elementwise affine, continuous regularization of an elementwise constant function by computing local averages

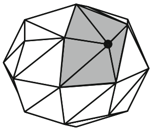


Fig. A.15 Nodal patch used to compute an average of a discontinuous function

and volume or surface area. In the case of an element, also the gradients of the nodal basis functions restricted to the element are provided.

A.4.3 Averaged Quantities

Given a discrete surface and an elementwise constant scalar quantity v_h, a continuous, elementwise affine approximation $\mathscr{A}_h[v_h]$ of v_h is computed in the routine `average_quant_surf.m` shown in Fig. A.14. The function $\mathscr{A}_h[v_h]$ is represented in the nodal basis by its nodal values that are defined by

$$\mathscr{A}_h[v_h](z) = \frac{\int_{\omega_z} v_h \, \mathrm{d}s}{\int_{\omega_z} 1 \, \mathrm{d}s},$$

for all nodes $z \in \mathscr{N}_h$ and with the support ω_z of the nodal basis function associated to z, cf. Figs. A.15.

A.4.4 Minors

For a list of matrices in $\mathbb{R}^{d \times d}$, $d = 2, 3$, that are identified with vectors in $\mathbb{R}^{d^2}$, the routine `minors.m` displayed in Fig. A.16 computes the nontrivial minors of the matrices, i.e., for a matrix $S \in \mathbb{R}^{d \times d}$, the determinant of S, if $d = 2$ and the vector

```
function val = minors(A)
if size(A,2) == 4
    val = A(:,1).*A(:,4)-A(:,2).*A(:,3);
elseif size(A,2) == 9
    val = zeros(size(A,1),10);
    val(:,1:9) = [A(:,5).*A(:,9)-A(:,6).*A(:,8),...
        A(:,4).*A(:,9)-A(:,6).*A(:,7),...
        A(:,4).*A(:,8)-A(:,5).*A(:,7),...
        A(:,2).*A(:,9)-A(:,3).*A(:,8),...
        A(:,1).*A(:,9)-A(:,3).*A(:,7),...
        A(:,1).*A(:,8)-A(:,2).*A(:,7),...
        A(:,2).*A(:,6)-A(:,3).*A(:,5),...
        A(:,1).*A(:,6)-A(:,3).*A(:,4),...
        A(:,1).*A(:,5)-A(:,2).*A(:,4)];
    val(:,10) = A(:,1).*val(:,1)-A(:,2).*val(:,2)...
        +A(:,3).*val(:,3);
end
```

Fig. A.16 Computation of the nontrivial minors of a list of $d \times d$ matrices that are identified with vectors in $\mathbb{R}^{d^2}$ for $d = 2, 3$

$$(\det S^{11}, \det S^{21}, \dots \det S^{33}, \det S)$$

if $d = 3$, where $S^{ij} \in \mathbb{R}^{2\times 2}$ denotes the matrix that is obtained by deleting the i-th row and j-th column of S.

A.4.5 Standard Finite Element Matrices

The routine `fe_matrices_weighted.m` shown in Fig. A.17 for elementwise constant functions $a_h, b_h : \Omega \to \mathbb{R}$ computes the $\mathcal{N}_h \times \mathcal{N}_h$ matrices with the entries

$$\int_\Omega a_h \nabla\varphi_z \cdot \nabla\varphi_y \,\mathrm{d}x, \quad \int_\Omega b_h \varphi_z \varphi_y \,\mathrm{d}x$$

for pairs of nodes $(z, y) \in \mathcal{N}_h \times \mathcal{N}_h$. For a triangle T with nodes z_1, z_2, z_3, a basis of the space of polynomials of maximal degree 2 on T is given by the functions

$$(\psi_1, \psi_2, \dots, \psi_6) = \big(\varphi_{z_1}, \varphi_{z_2}, \varphi_{z_3}, 4\varphi_{z_2}\varphi_{z_3}, 4\varphi_{z_3}\varphi_{z_1}, 4\varphi_{z_2}\varphi_{z_3}\big),$$

cf. Fig. A.18. The elementwise defined functions can be assembled to obtain a basis of the finite element space $\mathcal{S}^2(\mathcal{T}_h)$. The routine `fe_matrix_p2.m`, shown in Fig. A.19 computes the corresponding stiffness matrix.

```
function [s_a,m_b] = fe_matrices_weighted(c4n,n4e,a,b)
[nC,d] = size(c4n); nE = size(n4e,1);
m_loc = (ones(d+1,d+1)+eye(d+1))/((d+1)*(d+2));
ctr = 0; ctr_max = (d+1)^2*nE;
I = zeros(ctr_max,1); J = zeros(ctr_max,1);
X_s_a = zeros(ctr_max,1); X_m_b = zeros(ctr_max,1);
for j = 1:nE
    X_T = [ones(1,d+1);c4n(n4e(j,:),:)'];
    grads_T = X_T\[zeros(1,d);eye(d)];
    vol_T = det(X_T)/factorial(d);
    for m = 1:d+1
        for n = 1:d+1
            ctr = ctr+1;
            I(ctr) = n4e(j,m); J(ctr) = n4e(j,n);
            X_s_a(ctr) = vol_T*a(j)*grads_T(m,:)*grads_T(n,:)';
            X_m_b(ctr) = vol_T*b(j)*m_loc(m,n);
        end
    end
end
s_a = sparse(I,J,X_s_a,nC,nC); m_b = sparse(I,J,X_m_b,nC,nC);
```

Fig. A.17 Computation of weighted $P1$ finite element mass and stiffness matrix

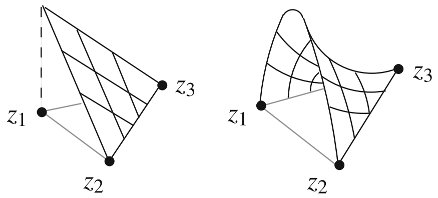


Fig. A.18 Typical hierarchical basis functions in the $P2$ finite element method with elementwise quadratic functions

A.4.6 Special Finite Element Matrices

The routine `nonlinear_fe_matrices_plast.m` shown in Fig. A.20 computes, for a stress function $\mathscr{S} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ and a given deformation $u_h : \Omega \to \mathbb{R}^d$, the vector

$$F(u_h)[e_p\varphi_z] = \int_\Omega \mathscr{S}(\nabla u_h) : \nabla(e_p\varphi_z)\,\mathrm{d}x - \int_\Omega f \cdot (e_p\varphi_z)\,\mathrm{d}x - \int_{\Gamma_\mathrm{N}} g \cdot (e_p\varphi_z)\,\mathrm{d}s$$

for $p = 1, 2, \ldots, d$ and $z \in \mathscr{N}_h$, the matrix

$$DF(u_h)[e_p\varphi_z, e_q\varphi_y] = \int_\Omega D\mathscr{S}(\nabla u_h)[\nabla(e_p\varphi_z), \nabla(e_q\varphi_y)]\,\mathrm{d}x$$

for $1 \le p, q \le d$ and $z, y \in \mathscr{N}_h$, and the stress field $\sigma_h = \mathscr{S}(\nabla u_h)$. The routine is a reduced version of the MATLAB program `nonlinear_fe_matrices.m`. The system matrix related to the discrete Kirchhoff element used for bending problems

```
function s_p2 = fe_matrix_p2(c4n,n4e,n4s,s4e,s_p1,vol_T)
nC = size(c4n,1); nE = size(n4e,1); nS = size(n4s,1);
s4e = s4e+nC;
m_loc = [2,1,1;1,2,1;1,1,2]/12;
shift = [2,1]; sides = [2,3;3,1;1,2];
s_p2 =  sparse(nC+nS,nC+nS); s_p2(1:nC,1:nC) = s_p1;
for j = 1:nE
    grads_T = [ones(1,3);c4n(n4e(j,:),:)']\[zeros(1,2);eye(2)];
    for k = 1:3
        for ell = 1:3
            for m = 1:2
                for n = 1:2
                  s_p2(s4e(j,k),s4e(j,ell)) = ...
                    s_p2(s4e(j,k),s4e(j,ell))+16*vol_T(j)...
                    *grads_T(sides(k,m),:)...
                    *grads_T(sides(ell,n),:)'...
                    *m_loc(sides(k,shift(m)),sides(ell,shift(n)));
                end
            end
        end
    end
    for k = 1:3
        for ell = 1:3
            for n = 1:2
                 s_p2(n4e(j,k),s4e(j,ell)) = ...
                   s_p2(n4e(j,k),s4e(j,ell))+4*(vol_T(j)/3)...
                   *grads_T(k,:)*grads_T(sides(ell,n),:)';
            end
            s_p2(s4e(j,ell),n4e(j,k)) = s_p2(n4e(j,k),s4e(j,ell));
        end
    end
end
```

Fig. A.19 Computation the $P2$ stiffness matrix

encodes the inner product

$$(v_h, w_h) \mapsto \int_\Omega \nabla\nabla_h v_h : \nabla\nabla_h w_h \, \mathrm{d}x$$

for all $v_h, w_h \in W_h$. The representing matrix with respect to an appropriate nodal basis of the finite element space $W_h \subset H^1(\Omega)$ is assembled in the MATLAB routine `fe_matrix_dkt.m` shown in Fig. A.21 using the $P2$ finite element stiffness matrix computed with the routine `fe_matrix_p2.m` displayed in Fig. A.19.

```
function [DF,F,sigma] = nonlinear_fe_matrices_plast(c4n,n4e,Nb,u)
[nC,d] = size(c4n); nE = size(n4e,1); nNb = size(Nb,1);
F = zeros(d*nC,1); sigma = zeros(nE,d^2); Du = zeros(nE,d^2);
ctr_max = d^2*(d+1)^2*nE; ctr = 0;
I = zeros(ctr_max,1); J = zeros(ctr_max,1); X = zeros(ctr_max,1);
for k = 1:d
    Du(:,(k-1)*d+(1:d)) = comp_gradient(c4n,n4e,u(k:d:nC*d));
end
for j = 1:nE
    [mp_T,vol_T,grads_T] = geometry_element(c4n(n4e(j,:),:));
    Du_T = Du(j,:); sigma(j,:) = stress(Du_T,j);
    for m = 1:d+1
        for p = 1:d
            D_psi_A = zeros(1,d^2);
            D_psi_A(d*(p-1)+(1:d)) = grads_T(m,:);
            D_Sigma_T_A = stress_derivative(Du_T,D_psi_A,j);
            for n = 1:d+1
                for q = 1:d
                    ctr = ctr+1;
                    D_psi_B = zeros(1,d^2);
                    D_psi_B(d*(q-1)+(1:d)) = grads_T(n,:);
                    I(ctr) = d*n4e(j,m)-d+p;
                    J(ctr) = d*n4e(j,n)-d+q;
                    X(ctr) = vol_T*D_Sigma_T_A*D_psi_B';
                end
            end
            phi_mp_T = zeros(d,1); phi_mp_T(p) = 1/(d+1);
            F(d*(n4e(j,m)-1)+p) = F(d*(n4e(j,m)-1)+p) ...
                +vol_T*(sigma(j,:)*D_psi_A'-f(mp_T)*phi_mp_T);
        end
    end
end
DF = sparse(I,J,X,d*nC,d*nC);
for j = 1:nNb
    [mp_S,vol_S] = geometry_side(c4n(Nb(j,:),:));
    for m = 1:d
        for p = 1:d
            phi_mp_S = zeros(d,1); phi_mp_S(p) = 1/d;
            F(d*(Nb(j,m)-1)+p) = F(d*(Nb(j,m)-1)+p)...
                -vol_S*g(mp_S)*phi_mp_S;
        end
    end
end

function val = f(x); global d; val = zeros(1,d);
function val = g(x); global d; val = zeros(1,d);
```

Fig. A.20 Vector, matrix, and stress field required in the solution of a nonlinear displacement problem; the nonlinear stress function and its derivative for a time step in elastoplasticity are provided by the routines `stress.m` and `stress_derivative.m`

```
function S_dkt = fe_matrix_dkt(c4n,n4e)
[n4s,s4e] = sides(n4e);
nC = size(c4n,1); nS = size(n4s,1);
D = sparse(2*(nC+nS),3*nC);
for j = 1:nC
    D(2*j-[1,0],3*j-[1,0]) = eye(2);
end
for j = 1:nS
    t_S = (c4n(n4s(j,2),:)-c4n(n4s(j,1),:))';
    ell_S = norm(t_S); t_S = t_S/ell_S;
    D(2*nC+2*j-[1,0],3*n4s(j,1)-2) = -3/(2*ell_S)*t_S;
    D(2*nC+2*j-[1,0],3*n4s(j,2)-2) = 3/(2*ell_S)*t_S;
    D(2*nC+2*j-[1,0],3*n4s(j,1)-[1,0]) = -(3/4)*(t_S*t_S');
    D(2*nC+2*j-[1,0],3*n4s(j,2)-[1,0]) = -(3/4)*(t_S*t_S');
end
[s_p1,¬,¬,vol_T] = fe_matrices(c4n,n4e);
s_p2 = fe_matrix_p2(c4n,n4e,n4s,s4e,s_p1,vol_T);
S = sparse(2*(nC+nS),2*(nC+nS));
S(1:2:2*(nC+nS),1:2:2*(nC+nS)) = s_p2;
S(2:2:2*(nC+nS),2:2:2*(nC+nS)) = s_p2;
S_dkt = D'*S*D;

function [n4s,s4e] = sides(n4e)
nE = size(n4e,1);
sides = reshape(n4e(:,[2,3,3,1,1,2])',2,[])';
[n4s,¬,sideNrs] = unique(sort(sides,2),'rows','first');
s4e = reshape(sideNrs(1:3*nE),3,[])';
```

Fig. A.21 System matrix for discrete Kirchhoff triangles

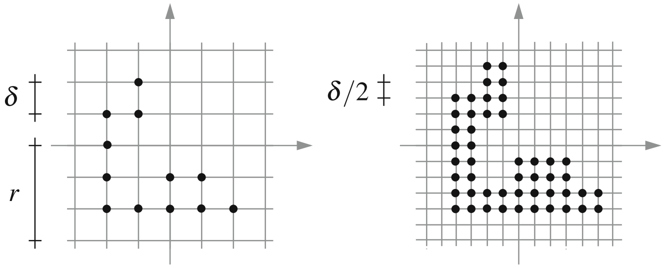


Fig. A.22 Subgrid $\mathcal{N}_{\delta,r} \subset \mathbb{R}^{d\times d}$ (*left*) and its local refinement obtained by adding atoms locally around existing ones (*right*)

A.5 Grids in $\mathbb{R}^{d\times d}$

A.5.1 Uniform Grids

The performance of MATLAB is suboptimal when large or iterated loops are required. This is the case in the generation of the grid $K_r^\infty \cap \delta\mathbb{Z}^{d\times d}$ in the d^2-dimensional space of matrices (Fig. A.22). To improve the performance this is realized in the C routine `grid_gen.c` shown in Fig. A.23. It employs the interface MEX that

```c
#include <math.h>
#include <mex.h>
void grid_2d(double atoms[], double delta, int M, int N, int L){
  int j, k, m, n, idx;
  for (j=-M; j<M+1; j++) for (k=-M; k<M+1; k++)
    for (m=-M; m<M+1; m++) for (n=-M; n<M+1; n++){
      idx = (j+M)*pow(N,3)+(k+M)*pow(N,2)+(m+M)*N+(n+M);
      atoms[0*L+idx] = j*delta; atoms[1*L+idx] = k*delta;
      atoms[2*L+idx] = m*delta; atoms[3*L+idx] = n*delta;}
}
void grid_3d(double atoms[], double delta, int M, int N, int L){
  int j, k, m, n, o, p, q, s, t, idx;
  for (j=-M; j<M+1; j++) for (k=-M; k<M+1; k++)
    for (m=-M; m<M+1; m++) for (n=-M; n<M+1; n++)
      for (o=-M; o<M+1; o++) for (p=-M; p<M+1; p++)
        for (q=-M; q<M+1; q++) for (s=-M; s<M+1; s++)
          for (t=-M; t<M+1; t++){
            idx = (j+M)*pow(N,8)+(k+M)*pow(N,7)+(m+M)*pow(N,6)
            +(n+M)*pow(N,5)+(o+M)*pow(N,4)+(p+M)*pow(N,3)
            +(q+M)*pow(N,2)+(s+M)*pow(N,1)+(t+M);
             atoms[0*L+idx] = j*delta; atoms[1*L+idx] = k*delta;
             atoms[2*L+idx] = m*delta; atoms[3*L+idx] = n*delta;
             atoms[4*L+idx] = o*delta; atoms[5*L+idx] = p*delta;
             atoms[6*L+idx] = q*delta; atoms[7*L+idx] = s*delta;
             atoms[8*L+idx] = t*delta;}
}
void mexFunction(int nlhs, mxArray *plhs[],
        int nrhs, const mxArray *prhs[]){
  double *delta, *r, *atoms;
  int d, L, M, N;
  if (nrhs!=3) mexErrMsgTxt("3 input arguments required!");
  if (nlhs!=1) mexErrMsgTxt("1 output argument required!");
  delta  = mxGetPr(prhs[0]);
  r = mxGetPr(prhs[1]);
  d = *mxGetPr(prhs[2]);
  M = floor((*r)/(*delta));
  N = 2*M+1;
  L = pow(N,pow(d,2));
  plhs[0] = mxCreateDoubleMatrix(L,pow(d,2),mxREAL);
  atoms =mxGetPr(plhs[0]);
  if (d==2)
    grid_2d(atoms,*delta,M,N,L);
  else
    grid_3d(atoms,*delta,M,N,L);
}
```

Fig. A.23 C routine grid_gen.c that generates the grid $K_r^{\infty} \cap \delta\mathbb{Z}^{d\times d}$; the program is incorporated into MATLAB with the interface MEX

```
#include <mex.h>
#include <math.h>
void new_grid_2d(double new_atoms[], double delta, double atoms[],
         int nr_old){
  int z, ctr, j, k, m, n, idx; idx = 16*nr_old;
  for (z=0; z<nr_old; z++){ctr = 0;
    for (j=0; j<2; j++) for (k=0; k<2; k++)
      for (m=0; m<2; m++) for (n=0; n<2; n++){
        new_atoms[0*idx+z*16+ctr] = atoms[0*nr_old+z]+j*delta;
        new_atoms[1*idx+z*16+ctr] = atoms[1*nr_old+z]+k*delta;
        new_atoms[2*idx+z*16+ctr] = atoms[2*nr_old+z]+m*delta;
        new_atoms[3*idx+z*16+ctr] = atoms[3*nr_old+z]+n*delta;
        ctr = ctr+1;}
}}
void new_grid_3d(double new_atoms[], double delta, double atoms[],
         int nr_old){
  int z,ctr, j, k, m, n, o, p, q, s, t, idx; idx = 512*nr_old;
  for (z=0;z<nr_old;z++){ctr = 0;
    for (j=0;j<2;j++) for (k=0;k<2;k++) for (m=0;m<2;m++)
      for (n=0;n<2;n++) for (o=0;o<2;o++) for (p=0;p<2;p++)
        for (q=0;q<2;q++) for (s=0;s<2;s++) for (t=0;t<2;t++){
          new_atoms[0*idx+z*512+ctr] = atoms[0*nr_old+z]+j*delta;
          new_atoms[1*idx+z*512+ctr] = atoms[1*nr_old+z]+k*delta;
          new_atoms[2*idx+z*512+ctr] = atoms[2*nr_old+z]+m*delta;
          new_atoms[3*idx+z*512+ctr] = atoms[3*nr_old+z]+n*delta;
          new_atoms[4*idx+z*512+ctr] = atoms[4*nr_old+z]+o*delta;
          new_atoms[5*idx+z*512+ctr] = atoms[5*nr_old+z]+p*delta;
          new_atoms[6*idx+z*512+ctr] = atoms[6*nr_old+z]+q*delta;
          new_atoms[7*idx+z*512+ctr] = atoms[7*nr_old+z]+s*delta;
          new_atoms[8*idx+z*512+ctr] = atoms[8*nr_old+z]+t*delta;
          ctr = ctr+1;}
}}
void mexFunction(int nlhs, mxArray *plhs[],
        int nrhs, const mxArray *prhs[]){
  double *delta, *atoms, *new_atoms;
  int d, m, d_sq;
  if (nrhs!=3) mexErrMsgTxt("3 input arguments required!");
  if (nlhs!=1) mexErrMsgTxt("1 output argument required!");
  m = mxGetM(prhs[1]);
  delta = mxGetPr(prhs[0]);
  atoms = mxGetPr(prhs[1]);
  d = *mxGetPr(prhs[2]); d_sq = pow(d,2);
  plhs[0] = mxCreateDoubleMatrix(pow(2,d_sq)*m,d_sq,mxREAL);
  new_atoms = mxGetPr(plhs[0]);
  if (d==2)
    new_grid_2d(new_atoms,*delta,atoms,m);
  else
    new_grid_3d(new_atoms,*delta,atoms,m);
}
```

Fig. A.24 C routine `loc_grid_ref.c` that locally adds new atoms to a given set of atoms

provides a simple way to incorporate C code in MATLAB. The routine is compiled under MATLAB with the command

```
mex grid_gen.c;
```

For this the gnu C compiler `gcc` has to be selected using the MATLAB command `mex -setup`. The routine can then be used within MATLAB with the command

```
atoms = grid_gen(delta,r,d);
```

The nodes of the grid are referred to as atoms.

A.5.2 Local Refinement

A grid or subgrid $\mathscr{N}_{\delta,r} \subset K_r^\infty \cap \delta\mathbb{Z}^{d\times d}$ can be refined locally by adding atoms in the neighborhoods of existing ones, i.e., by replacing every atom $s \in \mathscr{N}_{\delta,r}$ by the set of atoms

$$s + \frac{\delta}{2}\{0, 1\}^{d\times d},$$

cf. Fig. A.22 for a schematic description. This is implemented in the C program `loc_grid_ref.c` shown in Fig. A.24 that also employs the interface MEX. It is incorporated in MATLAB with the command

```
atoms_new = loc_grid_ref(delta/2,atoms,d);
```

Appendix B
Frequently Used Notation

Real numbers, vectors, and matrices

$\mathbb{Z}$	Integers
$\mathbb{N}$, $\mathbb{N}_0$	Positive and nonnegative integers
$\mathbb{R}$	Real numbers
$[s,t],\ (s,t)$	Closed and open interval
$\mathbb{R}^d$	d-dimensional Euclidean vector space
$\mathbb{R}^{n\times m}$	Vector space of n by m matrices
$B_r(x),\ B_r$	Open ball of radius r centered at x or at the origin
$K_r(x),\ K_r$	Closed ball of radius r centered at x or at the origin
$A \subset B$	A is a subset of B or $A = B$
$a,\ A$	(Column) vector and matrix
$a^\top,\ A^\top$	Transpose of a vector or matrix
$\lvert\cdot\rvert$	Euclidean length or Frobenius-norm
$a \cdot b = a^\top b$	Scalar product of vectors a and b
$a \otimes b = ab^\top$	Dyadic product of vectors a and b
$a \times b$	Cross product of vectors $a, b \in \mathbb{R}^3$
$a \perp b$	a is perpendicular to b
$A : B$	Inner product of matrices A and B
$\mathrm{tr} A$	Trace of the matrix A
I_L	$L \times L$ identity matrix
S^{m-1}	Unit sphere in $\mathbb{R}^m$
$SO(n)$	Group of special orthogonal matrices
$[x,y]^\top, (x,y)$	Vectors with entries x and y
$\begin{bmatrix} x_1 & x_2 \\ y_1 & y_2 \end{bmatrix}$	Matrix with entries x_1, x_2, y_1, y_2

S. Bartels, *Numerical Methods for Nonlinear Partial Differential Equations*,
Springer Series in Computational Mathematics 47,
DOI 10.1007/978-3-319-13797-1

Sets, domains, and functionals

Ω	Bounded Lipschitz domain in $\mathbb{R}^d$, $d = 2, 3$
n	Outer unit normal on $\partial\Omega$
Γ_{D}	Dirichlet boundary, closed subset of $\partial\Omega$
Γ_{N}	Neumann boundary, $\Gamma_{\mathrm{N}} = \partial\Omega \setminus \Gamma_{\mathrm{D}}$
$[0, T]$	Time interval
$\mathscr{A}$	Set of admissible functions or vector fields
I	Energy functional
W	Energy density

Linear spaces and operators

id	Identity operator
ker	Kernel of an operator
X, Y	Banach spaces
$\\|\cdot\\|_X$	Norm in X
X'	Linear bounded functionals $\Lambda : X \to \mathbb{R}$
$\langle\phi, x\rangle$	Duality pairing of $\phi \in X'$ and $x \in X$
$\\|\cdot\\|_{X'}$	Operator norm in X'
$\mathrm{L}(X, Y)$	Bounded linear operators $\Lambda : X \to Y$
$\\|\cdot\\|_{\mathrm{L}(X,Y)}$	Operator norm in $\mathrm{L}(X, Y)$
Λ'	Adjoint of $\Lambda \in \mathrm{L}(X, Y)$
H	Hilbert space
$(x, y)_H$	Inner product of x and y in a Hilbert space H

Differential operators

$\partial_i,\ \partial_{x_i},\ \frac{\partial}{\partial x_i}$	Partial derivative with respect to the i-th component
∇	Gradient of a function
div	Divergence of a vector field
D, D^2	Total derivative and Hessian of a function
$\partial_x,\ \partial_y,\ \partial_t,\ \partial^\alpha$	Partial derivatives
$\partial_n u$	Normal derivative $\nabla u \cdot n$ on $\partial\Omega$
$u_t,\ u'$	Partial derivative with respect to t
$\varepsilon(u)$	Symmetric gradient of a displacement
Δ	Laplace operator
δ	Fréchet derivative of a functional

Function spaces

$C^k(A;\mathbb{R}^m)$	k-times continuously differentiable vector fields
$C_c^\infty(\Omega;\mathbb{R}^m)$	Compactly supported, smooth vector fields
$C_0(\Omega;\mathbb{R}^m)$	Closure of $C_c^\infty(\Omega;\mathbb{R}^m)$ with respect to maximum norm
$L^p(\Omega;\mathbb{R}^m)$	Functions whose p-th power is Lebesgue integrable
$W^{k,p}(\Omega;\mathbb{R}^m)$	k-times weakly differentiable vector fields
$W_{\mathrm{D}}^{k,p}(\Omega;\mathbb{R}^m)$	Vector fields in $W^{k,p}(\Omega;\mathbb{R}^m)$ vanishing on Γ_{D} or $\partial\Omega$
$H^k(\Omega;\mathbb{R}^m)$	Sobolev space $W^{k,2}(\Omega;\mathbb{R}^m)$
$W^{k,p}([0,T];X)$	Sobolev-Bochner space of X-valued functions
$H_N(\mathrm{div};\Omega)$	Vector fields with square integrable divergence
$BV(\Omega)$	Functions of bounded variation
$SBV(\Omega)$	Special functions of bounded variation
$\\|\cdot\\|, (\cdot,\cdot)$	Norm and inner product in $L^2(\Omega;\mathbb{R}^m)$
$\|Du\|(\Omega)$	Total variation of the distributional derivative of u

Convex analysis

$\Gamma(H)$	Convex, proper, lower semicontinuous functionals on H
$\mathrm{dom}\,\psi$	Domain of the functional ψ
∂F	Subdifferential of $F \in \Gamma(H)$
F^*	Fenchel conjugate of F
I_C	Indicator functional of the convex set C
I	Convex functional
D	Dual of a convex functional I

Modes of convergence

$\to$	Strong convergence
$\rightharpoonup, \rightharpoonup^*$	Weak and weak* convergence
$\to^\Gamma$	Γ-convergence

Finite differences

τ	Step size
d_t	Backward difference quotient
$t_k,\ t_{k+1/2}$	Time steps $k\tau$ and $(k+1/2)\tau$
$u^k,\ u^{k+1/2}$	Approximations associated to time steps

Finite element spaces

$h,\ h_{\min}$	Maximal and minimal diameter of elements in $\mathscr{T}_h$
$h_T,\ h_S,\ h_z$	Local mesh-sizes
$\mathscr{N}$ or $\mathscr{N}_h$	Nodes that define vertices of elements
$\mathscr{T}$ or $\mathscr{T}_h$	Set of elements that define a triangulation
$\mathscr{S}$ or $\mathscr{S}_h$	Sides of elements in a triangulation
T or R	Element in a triangulation
$z,\ S$	Node, side
φ_z	Nodal basis function
ω_z	Patch of a node
$P_k(T)$	Polynomials of maximal degree k restricted to T
$\mathscr{L}^0(\mathscr{T}_h)$	$\mathscr{T}_h$-elementwise constant functions
$\mathscr{S}^1(\mathscr{T}_h)$	Continuous, $\mathscr{T}_h$-elementwise affine functions
$\mathscr{S}^1_{\mathrm{D}}(\mathscr{T}_h),\ \mathscr{S}^1_0(\mathscr{T}_h)$	Functions in $\mathscr{S}^1(\mathscr{T}_h)$ vanishing on $\varGamma_{\mathrm{D}}$ or $\partial\varOmega$
$\mathscr{I}$ or $\mathscr{I}_h$	Nodal interpolation operator on $\mathscr{T}_h$
$(v, w)_h$	Discrete inner product (mass lumping)
β_z	Integral of nodal basis function φ_z
$\mathscr{J}_h$	Clément quasi-interpolant
$[\![\nabla u_h \cdot n_S]\!]$	Jump of ∇u_h across S in direction of n_S
P_h	L^2-projection onto a discrete subspace
Q_h	H^1-projection onto a discrete subspace

Other notation

$c, C, C', C'', c_1, c_2, \ldots$	Mesh-size independent, generic constants
$\mathrm{d}x,\ \mathrm{d}s$	Volume and surface element for Lebesgue measure
$\mathrm{d}t$	Lebesgue integral with respect to time variable
$\overline{A}$	Closure of a set A
$\|A\|$	Volume or surface area of a set $A \subset \mathbb{R}^d$
$\operatorname{diam}(A)$	Diameter of the set A
χ_A	Characteristic function of a set A
δ_x	Dirac measure supported at x
δ_{ij}	Kronecker symbol
$\mathscr{O}(t), o(t)$	Landau symbols
$\operatorname{supp} f$	Support of a function f
$\mathscr{C}$	Consistency term
$\mathscr{R}$	Residual functional

MATLAB **routines**

`d`	Space dimension
`red`	Number of uniform refinements
`c4n`	List of coordinates of nodes
`n4e`	List of elements
`Db`, `Nb`	Lists of sides on Γ_{D} and Γ_{N}
`dNodes`	Nodes belonging to Γ_{D}
`fNodes`	Nodes not belonging to Γ_{D}
`nC`, `nE`, `nDb`, `nNb`	Number of nodes, elements and sides on Γ_{D} and $\overline{\Gamma}_{\mathrm{N}}$
`s`, `m`, `m_lumped`	$P1$ stiffness, mass, and lumped mass matrix
`m_Nb`, `m_Nb_lumped`	Exact and discrete inner products on Γ_{N}
`vol_T`	Areas or volumes of elements
`grads_T`	Elementwise gradients of nodal basis functions
`mp_T`	Midpoints of elements
`tau`	Step size
`I`, `J`, `X`	Lists to generate a sparse matrix

Index

S. Bartels, *Numerical Methods for Nonlinear Partial Differential Equations*,
Springer Series in Computational Mathematics 47,
DOI 10.1007/978-3-319-13797-1