Adaptive Monotone Multigrid Methods for Nonlinear Variational Problems

R. Kornhuber

Preface

This work is dedicated to the memory of Heinrich and Marie Barnstorf

A wide range of problems occurring in engineering and industry is characterized by the presence of a free (i.e. a priori unknown) boundary where the underlying physical situation is changing in a discontinuous way. Mathematically, such phenomena can be often reformulated as variational inequalities or related non–smooth minimization problems.

In these research notes, we will describe a new and promising way of constructing fast solvers for the corresponding discretized problems providing globally convergent iterative schemes with (asymptotic) multigrid convergence speed. The presentation covers physical modelling, existence and uniqueness results, finite element approximation and adaptive mesh–refinement based on a posteriori error estimation. The numerical properties of the resulting adaptive multilevel algorithm are illustrated by typical applications, such as semiconductor device simulation or continuous casting.

Essential parts of these notes were completed during the last of 7 years which I spent as a researcher at the Konrad–Zuse–Center in Berlin (ZIB). It is a pleasure for me to express my deep gratitude to P. Deuflhard, president of ZIB, for continuous encouragement and support. He directed my interest to the field of adaptive multilevel methods, taught me the mutual benefits of theoretical research and practical applications, and created a stimulating, productive atmosphere in the whole institute. Many ideas to be presented here are heavily influenced by the discussions and collaborations with my former colleagues at the ZIB, in particular with F. Bornemann and R. Roitzsch. Special thanks to R. Beck and B. Erdmann for invaluable computational assistance.

The preparation of a final version of these notes was supported by various important remarks and suggestions of my former colleagues at the Weierstraß–Institute Berlin and by the careful proofreading of Mrs. P. Lawday from the ICA Stuttgart.

I am obliged to R.H.W. Hoppe for friendly collaboration over more than a decade, to H. Yserentant and G. Wittum for their continuous encouragement, and to R.D. Grigorieff, who laid the foundations of my mathematical thinking.

Finally, I want to thank Martina, Kai and Marja for patiently overlooking my mental absence for quiet a while.

Stuttgart, June 1996 Ralf Kornhuber

Contents

Introduction

Since the pioneering papers of Fichera [55] and Stampacchia [113] in the early sixties, variational inequalities have proved extremely useful for the formulation of a wide range of problems from mechanics, physical and biological science, metallurgy, soil mechanics, etc. A characteristic feature of most of these applications is the presence of a *free* or *moving boundary* dividing the computational domain into different *phases* with different physical properties.

An important subclass of such problems can be rewritten in the form

$$\mathcal{J}(u) + \phi(u) = \text{ min.}$$

Obstacle problems or time–discretized two–phase Stefan problems are typical examples. While the usual quadratic energy functional $\mathcal{J}$ has all the nice properties of a parabola, the additional convex functional ϕ defined by

$$\phi(v) = \int_\Omega \Phi(v(x))\, dx$$

has only some of them. In particular, ϕ does not need to be differentiable. A free boundary separates the regions in which $\Phi(u)$ is smooth.

In these notes, we will try to work out some of the consequences originating from this additional functional ϕ. Our reasoning will be guided by the questions whether there is a solution, how we can compute it, and what it is good for. This pragmatic approach will lead us from a bit of physical modelling to the roots of functional analysis, from the convergence of finite element approximations to the convergence rates of multigrid methods, and, finally, to some practical computations which will leave us with more questions than answers.

In the beginning, the reader should be familiar with elementary facts on Hilbert spaces. Some basic notions of convex analysis will be introduced in the first chapter, where we summarize well–known existence and uniqueness

results and outline the convergence analysis of a finite element discretization. In doing so, we never strive for generality, but for a clear presentation of basic concepts. However, this introductory chapter could be used as a starting point for a more advanced course in nonlinear functional analysis or optimization.

The unified view on multigrid methods and domain (or space) decomposition has caused a breakthrough in the understanding of adaptive multilevel methods for selfadjoint elliptic problems. We recommend the fundamental monograph of Hackbusch [65] and refer to Bramble [31], Dryja and Widlund [47], Xu [122] or Yserentant [126] for recent developments. With this background, the main part of these notes is devoted to the construction of fast solvers for the finite element analogue of our continuous minimization problem.

Standard relaxation methods (cf. e.g. Glowinski [61]) are globally convergent, but usually suffer from rapidly deteriorating convergence rates with increasing refinement. We first explain how to incorporate a certain redundancy in these methods which is intended to increase the convergence speed. It turns out that *monotonically decreasing energy* $\mathcal{J} + \phi$ is crucial for the global convergence of the resulting *extended relaxation schemes*. This property is preserved by suitable perturbations and will be the essential feature of *monotone multigrid methods*.

Constructing a multigrid method for a (discrete) free boundary problem, one always has to answer the question how to represent the free boundary on the coarse grids. Our answer is that there is no such representation. As a consequence, the coarse grid correction must not change the actual guess of the free boundary (resulting from fine grid smoothing). This condition applies *locally* to each correction from each coarse grid node and, for theoretical purposes, can be regarded as a *local damping* of the coarse grid correction. However, the invariance of the actual free boundary may exclude a large number of coarse grid nodes from contributing to the correction and this may again deteriorate the speed of convergence. A possible remedy is to adapt the underlying space decomposition to the actual free boundary. In this way, we obtain so–called *truncated monotone multigrid methods* in contrast to the *standard version* introduced above. For both methods, we will derive upper bounds for the asymptotic convergence rates depending only logarithmically on the minimal stepsize. The existence of related global bounds is still an open question.

In the special case of obstacle problems, our standard monotone multigrid method reduces to the algorithm of Mandel [92, 93] and the truncated variant can be regarded as a further development of well–known heuristic approaches (cf. Brandt and Cryer [33], Hackbusch and Mittelmann [66]), see Kornhuber [82] for details. Other multigrid methods have been derived for other special cases, such as semi-discrete Stefan problems (see e.g. Hoppe [72] or Hoppe and Kornhuber [73]). For a comparison, we refer to Kornhuber [83].

There should be no confusion with the monotone multigrid methods introduced by Zou [129] and improved by other authors (cf. Bäsler and Törnig [16] and Voller [120]). These methods are designed for nonlinear systems involving (generalized) M-functions and provide converging sequences of sub– and supersolutions.

In order to provide a hierarchy of grids together with an efficient finite element approximation, the underlying triangulation should be selected adaptively on the basis of efficient and reliable a posteriori error estimates. As we are dealing with a minimization problem, it is natural to control the error in the corresponding energy norm. Then, *hierarchical error estimates* provide a unifying framework for the construction of a posteriori estimates for linear selfadjoint problems (cf. Deuflhard, Leinen and Yserentant [45], Bank and Smith [10], Bornemann, Erdmann, and Kornhuber [28]). Chapter 4 contains some steps towards the extension of this concept to the non–smooth minimization problem under consideration.

In the final chapter, finite element discretization, iterative solution and a posteriori error estimation are assembled to an *adaptive multilevel method*. Using this algorithm as some sort of black box method, we consider four examples of different nature, ranging from pure model problems to quite realistic situations. In all our experiments, we observed a similar efficiency and reliability of our adaptive multilevel algorithm as for related elliptic selfadjoint problems. Nevertheless, the results of the numerical computations are in turn a source of challenging theoretical and algorithmical questions which we hope will stimulate future research.

1 Nonlinear Variational Problems

In this introductory chapter, we briefly address the mathematical modelling, the mathematical analysis and the discretization of certain free boundary problems. Though all material to be presented is well–known from several monographs, we mostly include the proofs. One reason is to make the presentation as self–contained as possible and the other is to illustrate the intimate relation of mathematical physics, functional analysis and, last but not least, numerics.

1.1 Free and Moving Boundary Problems

Non–smooth minimization problems, elliptic variational inequalities and variational inclusions have proved to be essential in a wide range of problems from physics and engineering, particularly those with a *free boundary*. For an overview, we refer to the monographs of Crank [41], Duvaut and Lions [48], Elliott and Ockendon [51], Friedman [57], Friedman and Spruck [58], Glowinski [61], Glowinski, Lions and Trémolières [62], Hlaváček, Haslinger, Nečas and Lovíšek [68], Rodrigues [105] and the literature cited therein.

In order to motivate the abstract setting of our subsequent considerations, we will now take a closer look at three typical examples: an obstacle problem, a time–discretized Stefan problem and a time–discretized porous medium equation.

The latter two examples originate from the application of Rothe's method (cf. Rothe [108]) to degenerate parabolic problems. Basic existence, uniqueness, and regularity results for the Stefan problem are presented in the monographs of Rubinstein [109], Jerome [78] and Meirmanov [96], see also the earlier work of Oleinik [101], Kamenomostskaya [79] and Friedman [56]. Important contributions to the porous medium equation were made by Oleinik et al. [102], Caffarelli and Friedman [37], Alt and Luckhaus [2] and others. We refer to the surveys of Aronson [3] and Vazquez [116] for further information. Optimal error estimates for the semi–discretization in

time were given very recently by Rulla and Walkington [110], improving previous results of Jerome [78].

Rothe's method is not only a well–established way of showing existence and uniqueness [78] but also provides a powerful numerical approach. In the light of the fundamental work of Bornemann [23, 24, 25] on the linear parabolic case, the efficient numerical treatment of the spatial problems is a crucial step towards a fast adaptive algorithm for the original time–dependent problem.

1.1.1 Deformation of a Membrane with a Rigid Obstacle

In classical elasticity theory, a membrane is a thin plate offering no resistance to bending. Let us consider a membrane with uniform tension attached to the boundary $\partial\Omega$ of a domain $\Omega \subset \mathbb{R}^2$. The membrane is subjected to the action of a vertical force with density f and, in addition, must lie below a given rigid obstacle with height φ. We are interested in the vertical displacement u corresponding to the equilibrium position. A one–dimensional analogue is shown in Figure 1.1.

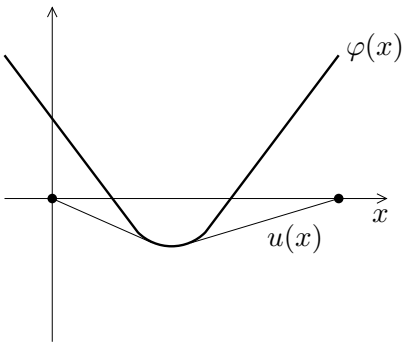

Figure 1.1 Membrane with upper obstacle φ

It is clear that the above conditions can be written as

$$u \le \varphi \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega. \tag{1.1}$$

Among all admissible states of the membrane, the equilibrium is attained at the state with *minimal total energy*. For a given displacement v, the total

energy $\mathcal{J}(v)$ consists of two contributions reflecting the *tension* and the *displacement* of the membrane respectively.

The contribution $\mathcal{J}_1(v)$ associated with the tension is proportional to the increase of area resulting from the deformation,

$$\mathcal{J}_1(v) = \alpha \int_\Omega (\sqrt{1 + v_{x_1}^2 + v_{x_2}^2} - 1)\, dx, \quad \alpha > 0. \tag{1.2}$$

Limiting our considerations to small strains, we neglect the higher–order terms in (1.2) to obtain

$$\mathcal{J}_1(v) \doteq \tfrac{1}{2} \int_\Omega \alpha |\nabla v|^2\, dx. \tag{1.3}$$

The second contribution $\mathcal{J}_2(v)$ associated with the displacement (subject to the force density f) is given by

$$\mathcal{J}_2(v) = -\int_\Omega f v\, dx. \tag{1.4}$$

In the light of (1.3) and (1.4), the total energy $\mathcal{J}(v)$ turns out to be

$$\mathcal{J}(v) = \tfrac{1}{2} a(v, v) - \ell(v), \tag{1.5}$$

where the bilinear form $a(\cdot, \cdot)$ and the linear functional ℓ are defined by

$$a(v, w) = \int_\Omega \alpha \nabla v \cdot \nabla w\, dx, \qquad \ell(v) = \int_\Omega f v\, dx. \tag{1.6}$$

The set $\mathcal{K}$ of admissible displacements

$$\mathcal{K} = \{v \in H^1(\Omega) \mid v \le \varphi \text{ a.e. in } \Omega,\ v = 0 \text{ on } \partial\Omega\} \tag{1.7}$$

consists of all v with finite energy satisfying the conditions (1.1). Assuming that φ is smooth enough (i.e. $\varphi \in H^1(\Omega)$) and non–negative on $\partial\Omega$, we will see later on that $\mathcal{K}$ is a non-empty, closed, convex subset of $H_0^1(\Omega)$.

Finally, the displacement u representing the equilibrium position of the membrane must be a solution of the following minimization problem

$$u \in \mathcal{K}: \quad \mathcal{J}(u) \le \mathcal{J}(v), \quad \forall v \in \mathcal{K}. \tag{1.8}$$

Note that (1.8) can be regarded as an extension of the classical Dirichlet principle. Indeed, if no obstacle is present, we clearly have $\mathcal{K} = H_0^1(\Omega)$ and u satisfies Euler's equation associated with (1.8) which turns out to be the weak form of Poisson's equation $\Delta u = f/\alpha$ with homogeneous Dirichlet boundary conditions.

1.1.2 A Semi–Discrete Stefan Problem

We consider the melting and solidification of some stationary substance occupying the domain $\Omega \subset \mathbb{R}^2$ during the time interval $[0,T]$. We assume that the phase change takes place at the fixed temperature θ_*. Then the temperature θ satisfies $\theta > \theta_*$ in the liquid fraction $\Omega_+(t)$ and $\theta < \theta_*$ in the solid fraction $\Omega_-(t)$, $t \in [0,T]$, respectively. Both phases are separated by a free boundary $\Gamma(t)$, $t \in [0,T]$. We assume for the moment that Γ is a smooth manifold with normal n_Γ pointing in the outward direction of Ω_+ (see Figure 1.2).

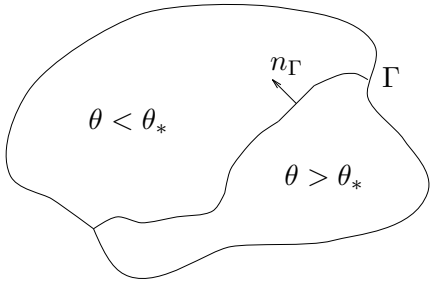

Figure 1.2 Heat flow with phase transition

Let $\mathcal{E}$ denote the *specific internal energy* or *heat content*, $\vec{v}$ describes the *heat flux* and F stands for a body heating term. Then the *conservation of energy* implies that

$$\rho \frac{\partial}{\partial t} \int_{\Omega'} \mathcal{E}(x,t)\, dx + \int_{\partial\Omega'} \vec{v}(x,t) \cdot n'\, d\sigma = \int_{\Omega'} F\, dx \tag{1.9}$$

holds for each fixed subset $\Omega' \subset \Omega$ with outward normal n'. For simplicity, we only consider *constant density* ρ. Selecting subsets $\Omega' \subset \Omega_+$ or $\Omega' \subset \Omega_-$ and assuming sufficient regularity, we can apply the divergence theorem to derive the pointwise equation

$$\rho \frac{\partial}{\partial t} \mathcal{E} + \nabla \cdot \vec{v} = F \quad \text{in } \Omega_+ \cup \Omega_-. \tag{1.10}$$

If Ω' intersects the free boundary $\Gamma(t)$, then the derivation in time no longer commutes with the integration and the divergence theorem has to be applied

separately in the two phases. The resulting boundary terms give rise to the well–known *Stefan condition* (cf. Stefan [114])

$$\rho[\mathcal{E}]_-^+ V_\Gamma = [\vec{v}]_-^+ n_\Gamma \quad \text{on } \Gamma. \tag{1.11}$$

The Stefan condition relates the velocity V_Γ of the free boundary in normal direction n_Γ with the jumps of $\mathcal{E}$ and $\vec{v}$ at the interface. The jumps are defined by

$$[w]_-^+ = \lim_{x\to\Gamma, x\in\Omega_+} w(x) - \lim_{x\to\Gamma, x\in\Omega_-} w(x). \tag{1.12}$$

In order to express the heat content $\mathcal{E}$ in terms of the temperature θ, we assume the thermodynamic relation

$$\mathcal{E}(\theta) = \int_{\theta_*}^{\theta} c(\vartheta)\, d\vartheta + s(\theta), \quad s(\theta) = \begin{cases} 0 & \text{if } \theta < \theta_* \\ [0,L] & \text{if } \theta = \theta_* \\ L & \text{if } \theta > \theta_* \end{cases}, \tag{1.13}$$

introducing the *heat capacity* $c(\theta)$ and the *latent heat* $L > 0$. Note that the *enthalpy function* $\mathcal{E}(\theta)$ is set–valued at the phase change temperature θ_*.

The heat flux $\vec{v}$ is given by *Fourier's law*

$$\vec{v} = -\kappa(\theta)\nabla\theta, \tag{1.14}$$

where $\kappa(\theta)$ denotes the *thermal conductivity*.

We assume that the heat capacity $c(\theta)$ and the thermal conductivity $\kappa(\theta)$ have positive constant values c_+, c_- and κ_+, κ_- over the two phases Ω_+, Ω_-, respectively.

Inserting (1.13) and (1.14) in (1.10) and (1.11), we obtain the *classical formulation of the two–phase Stefan problem*

$$\begin{aligned} \rho c(\theta)\theta_t - \nabla(\kappa(\theta)\nabla(\theta)) &= F && \text{in } \Omega_+ \cup \Omega_- \\ \theta = \theta_*, \qquad \rho L V_\Gamma &= [\kappa(\theta)\nabla\theta]_-^+ n_\Gamma && \text{on } \Gamma \end{aligned} \tag{1.15}$$

which has to be completed by suitable initial and boundary conditions. As a rule, either the temperature or the heat flux is prescribed at the boundary of Ω and the temperature is assumed to be given at the initial time instant.

The classical formulation (1.15) is extended to distributions in $\mathcal{D}'(Q)$, $Q = \Omega \times (0,T)$ by using generalized derivatives in space and time. Additionally, we drop the assumption that the free boundary must be a smooth manifold. Recall that $\mathcal{E}(\theta)$ is set–valued on the transition zone Γ which now may have non–zero measure in $\mathbb{R}^2$. By definition, θ is a *generalized solution of the two–phase Stefan problem*, if

$$\rho \frac{\partial}{\partial t} W - \nabla(\kappa(\theta)\nabla\theta) = F, \quad W \in \mathcal{E}(\theta), \tag{1.16}$$

holds in the sense of distributions in $\mathcal{D}'(Q)$.

The generalized formulation (1.16) of the two–phase Stefan problem contains (1.15) as a special case. In particular, a solution θ of (1.16) which is sufficiently regular and admits a classical smooth free boundary satisfies the Stefan condition (1.11). This follows from Green's formula and the representation of the normal velocity $V_\Gamma = N_t/\|N_\Gamma\|$ by the normal vector $N_\Sigma = (N_\Gamma, N_t)$ on the 2-dimensional manifold $\Sigma = \{(x,t) \mid \theta(x,t) = \theta_*\} \subset Q$ oriented in the direction of the solid phase.

Introducing a normalized temperature $U = K(\theta)$ and a normalized enthalpy $\mathcal{H}(U) = \rho\mathcal{E}(K^{-1}(U))$ via the standard Kirchhoff transformation

$$U = K(\theta) = \int_{\theta_*}^{\theta} \kappa(\vartheta)\, d\vartheta = \begin{cases} \kappa_-(\theta - \theta_*) & \text{if } \theta \le \theta_* \\ \kappa_+(\theta - \theta_*) & \text{if } \theta \ge \theta_* \end{cases},$$

we transform the doubly nonlinear problem (1.16) into a differential inclusion of the form

$$\frac{\partial}{\partial t} W - \Delta U = F, \quad W \in \mathcal{H}(U), \tag{1.17}$$

where the normalized enthalpy $\mathcal{H}$ becomes

$$\mathcal{H}(U) = \begin{cases} \rho\frac{c_-}{\kappa_-}U & \text{if } U < 0 \\ [0, \rho L] & \text{if } U = 0 \\ \rho\frac{c_+}{\kappa_+}U + L & \text{if } U > 0 \end{cases}. \tag{1.18}$$

Observe that we have to prescribe boundary conditions and an initial enthalpy $\mathcal{H}|_{t=0}$ rather than an initial temperature to ensure uniqueness.

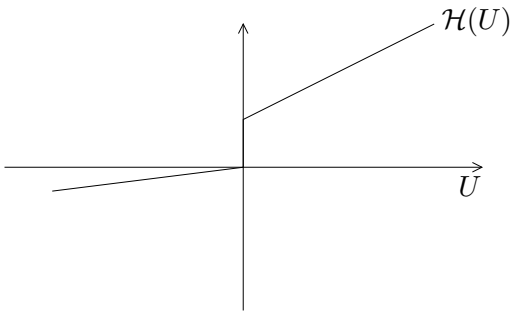

Figure 1.3 The normalized enthalpy $\mathcal{H}$

We now discretize (1.17) in time, using the backward Euler scheme with respect to a given grid

$$0 = t_0 < t_1 < \ldots < t_I = T \tag{1.19}$$

with step size $\tau_i = t_i - t_{i-1}$. Then we have to solve an *elliptic differential inclusion*

$$\tau_i F(\cdot, t_i) + W_{i-1} + \tau_i \Delta U_i \in \mathcal{H}(U_i) \tag{1.20}$$

in each time step. The unknown U_i is an approximation of $U(\cdot, t_i)$ and $W_{i-1} \in \mathcal{H}(U_{i-1})$ is a selection of the enthalpy from the preceding time step. For simplicity, we only consider homogeneous Dirichlet boundary conditions. Using the bilinear form $a(\cdot, \cdot)$ and the linear functional ℓ defined by

$$a(v, w) = \int_\Omega \tau_i \nabla v \cdot \nabla w dx, \quad \ell(v) = \int_\Omega f v \, dx, \tag{1.21}$$

with $f = \tau_i F(\cdot, t_i) + W_{i-1}$, a weak form of the spatial problem (1.20) is given by the *elliptic variational inclusion*

$$u \in H: \quad \ell(v) - a(u, v) \in (\mathcal{H}(u), v)_{L^2(\Omega)}, \quad \forall v \in H, \tag{1.22}$$

where we have set $u = U_i$, $H = H_0^1(\Omega)$, and

$$(\mathcal{H}(u), v)_{L^2(\Omega)} = \{(w, v)_{L^2(\Omega)} \mid w \in \mathcal{H}(u) \text{ a.e. in } \Omega\}. \tag{1.23}$$

1.1.3 The Semi–Discrete Porous Medium Equation

We consider a homogeneous gas flowing through a homogeneous porous medium occupying a domain $\Omega \subset \mathbb{R}^2$ during the time interval $[0,T]$. We assume that $\Omega_+(t) \subset \Omega$ is saturated with gas at the time t, while no gas is present in the remaining part of Ω. The (free) boundary $\Gamma(t)$ of $\Omega_+(t)$, $t \in [0,T]$, is supposed to be smooth and has the outward normal n_Γ (cf. Figure 1.4).

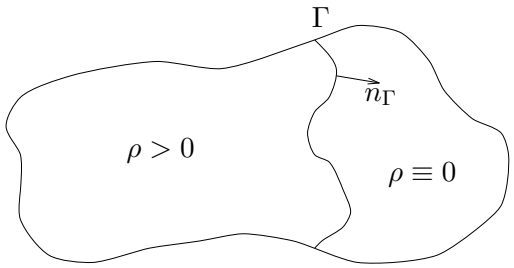

Figure 1.4 Gas flow through a porous medium

The *conservation of mass* implies that

$$\frac{\partial}{\partial t} \int_{\Omega'} \vartheta \rho(x,t)\, dx + \int_{\partial\Omega'} \rho(x,t) \vec{v}(x,t) \cdot n' d\sigma = 0 \tag{1.24}$$

holds for each subset $\Omega' \subset \Omega$ with outward normal n'. The *porosity* ϑ represents the portion of the area of Ω' which is available for the flow and ρ is the *density* of the gas. Assuming sufficient regularity, (1.24) can be interpreted as the differential equation (cf. (1.10))

$$\frac{\partial}{\partial t}(\vartheta\rho) + \nabla(\rho\vec{v}) = 0 \quad \text{in } \Omega_+. \tag{1.25}$$

Observe that the mass flux $\rho\vec{v}$ is continuous across Γ.

We suppose that the flow is governed by *Darcy's law* (cf. Darcy [42])

$$\vec{v} = -\tfrac{k}{\mu}\, \nabla p, \tag{1.26}$$

where k is the *capillarity* of the porous medium, μ is the *viscosity* and p is the *pressure* of the gas.

As both the porous medium and the gas are homogeneous, ϑ, k, and μ are positive constants. In order to relate the density to the pressure, we suppose that the equation of state

$$\rho = (p_0\, p)^{1/(m-1)}, \quad p \geq 0, \tag{1.27}$$

holds with real constants $p_0 > 0$, $m \geq 2$. If the flow is isothermic, then $m = 2$, while $m > 2$ holds for an adiabatic process.

Inserting (1.26) and (1.27) in (1.25), we obtain the *classical formulation of the porous medium equation*

$$\begin{aligned} &\frac{\partial}{\partial t}\rho = \Delta(\beta\rho^m) && \text{in } \Omega_+ \\ &\rho = 0, \qquad [\beta\nabla\rho^m]^+_- n_\Gamma = 0 && \text{on } \Gamma \end{aligned} \tag{1.28}$$

denoting $\beta = k(m-1)/\mu\rho_0 m > 0$. The jump $[\beta\nabla\rho^m]^+_-$ of the (scaled) mass flux across Γ is defined according to (1.12). For $m = 2$, (1.28) is known as the Boussinesq equation modelling for example the unsteady flow in a phreatic aquifer (see e.g. Bear [18]).

In addition, we have to prescribe the initial density and suitable boundary conditions for $t > 0$. The resulting initial boundary value problem is parabolic for $\rho > 0$ but degenerates when $\rho = 0$. The most striking manifestation of the degeneracy of this equation is that the free boundary Γ is propagating with finite speed.

Before we derive a weak formulation of (1.28), we formally extend (1.28) to negative densities by substituting ρ^m by ρ_+^m,

$$\rho_+ = \max\{0, \rho\}.$$

Other extensions are possible (cf. [2, 78]). The resulting differential equation

$$\frac{\partial}{\partial t}\rho = \Delta(\beta\rho_+^m) \tag{1.29}$$

has to be understood in the sense of distributions on $\mathcal{D}'(Q)$, $Q = \Omega \times (0,T)$. In this way, the interface conditions appearing in (1.28) are implicitly incorporated in (1.29).

In analogy to the preceding section, we introduce the Kirchhoff–type transformation

$$U = K(\rho) = \beta \rho_+^m \qquad K^{-1}(U) = \begin{cases} \frac{1}{\beta} \sqrt[m]{U} & \text{if } U > 0 \\ [0, -\infty) & \text{if } U = 0 \end{cases}.$$

Denoting $\mathcal{P}(U) = K^{-1}(U)$, we can rewrite the differential equation (1.29) as the differential inclusion

$$\frac{\partial}{\partial t} W = \Delta U, \qquad W \in \mathcal{P}(U). \tag{1.30}$$

We emphasize that the degeneracy of the problem is now represented by the fact that $\mathcal{P}$ is not Lipschitz.

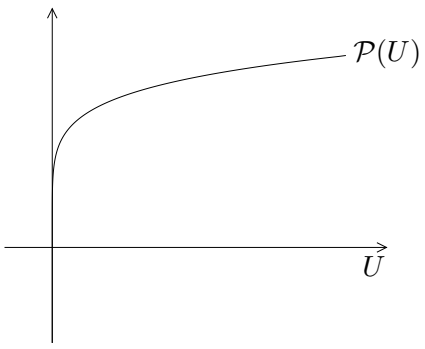

Figure 1.5 The inverse Kirchhoff transformation $\mathcal{P}$

The implicit time discretization by the backward Euler scheme with respect to a given grid (1.19) with step size $\tau_i = t_i - t_{i-1}$ leads to the spatial problems

$$W_{i-1} + \tau_i \Delta U_i \in \mathcal{P}(U_i), \tag{1.31}$$

where $W_{i-1} \in \mathcal{P}(U_{i-1})$ is an appropriate selection and U_i approximates $U(\cdot, t_i)$, $i = 1, \ldots, I$. Observe that discretizing (1.29) directly in time would lead to *quasilinear* spatial problems instead of the *semilinear* elliptic differential inclusions (1.31).

A weak form of (1.31) is given by the *elliptic variational inclusion*

$$u \in H: \quad \ell(v) - a(u,v) \in (\mathcal{P}(u), v)_{L^2(\Omega)}, \quad \forall v \in H, \tag{1.32}$$

where we have set $u = U_i$ and $H = H_0^1(\Omega)$ incorporates homogeneous Dirichlet boundary conditions. The bilinear form $a(\cdot,\cdot)$ and the linear functional ℓ are taken from (1.21) and the set $(\mathcal{P}(u), v)_{L^2(\Omega)}$ is defined in analogy to (1.23).

Observe that the semi–discrete porous medium equation (1.32) and the semi–discrete two–phase Stefan problem (1.22) formally coincide. However, in (1.32), the set–valued function $\mathcal{P}$ is no longer piecewise linear but piecewise smooth.

1.2 Convex Minimization

In this section, we introduce a non–smooth minimization problem which will turn out to contain the three examples given above as special cases. After a precise definition of the problem and a discussion of the basic assumptions, we consider the existence and uniqueness of solutions. For this reason, we give a brief introduction to convex analysis, presenting only the very basic results which are needed here. For further reading, we refer to the monographs of Aubin [4], Clarke [40], Deimling [43] or Ekeland and Temam [49]. Applications to elliptic variational inclusions can be found in the work of Barbu [15], Brézis [34, 35] or Jerome [78]. We do not address the question of regularity of solutions, but recommend the monographs of Baiocchi and Capelo [7], Kinderlehrer and Stampacchia [81], Rodriguez [105] and the literature cited therein.

1.2.1 The Continuous Problem

Let Ω be a bounded, polygonal domain in the Euclidean space $\mathbb{R}^2$. If a result cannot be generalized to three dimensions, this will be pointed out explicitely. We consider the *minimization problem*

$$u \in H: \quad \mathcal{J}(u) + \phi(u) \le \mathcal{J}(v) + \phi(v), \quad \forall v \in H, \tag{2.33}$$

on a closed subspace $H \subset H^1(\Omega)$. The quadratic functional $\mathcal{J}$,

$$\mathcal{J}(v) = \tfrac{1}{2}a(v,v) - \ell(v), \tag{2.34}$$

is induced by a continuous, symmetric and H–elliptic bilinear form $a(\cdot,\cdot)$ and a bounded, linear functional ℓ on H. Recall that $a(\cdot,\cdot)$ is *H–elliptic* if

$$\alpha\|v\|^2_{H^1(\Omega)} \le a(v,v), \quad \forall v \in H, \tag{2.35}$$

holds with a generic constant $\alpha > 0$. By virtue of the assumptions on the bilinear form $a(\cdot,\cdot)$, the *energy norm*

$$\|v\| = a(v,v)^{1/2} \tag{2.36}$$

is equivalent to the usual Sobolev norm on the solution space H. For simplicity, we select $H = H^1_0(\Omega)$ corresponding to homogeneous Dirichlet boundary conditions. Other boundary conditions of Neumann or mixed type can be treated in the usual way.

The functional ϕ has the form

$$\phi(v) = \int_\Omega \Phi(v(x))\,dx. \tag{2.37}$$

We impose the following conditions on the scalar function Φ.

(V1) $\Phi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is convex, i.e.

$$\Phi(\omega z + (1-\omega)z') \le \omega\Phi(z) + (1-\omega)\Phi(z'), \quad \forall \omega \in [0,1], \quad \forall z, z' \in \mathbb{R}.$$

(V2) $K = \{z \in \mathbb{R} \mid \Phi(z) < +\infty\}$ is a closed interval with $0 \in K$.

The subset $K = \operatorname{dom}\Phi$ is the *domain* of Φ. It follows from the convexity that Φ is locally Lipschitz on the interior of K (see e.g. Aubin [4]). We require the stronger condition that

(V3) $$|\Phi(z) - \Phi(z')| \le G(|z| + |z'|)|z - z'|, \quad z, z' \in K,$$

holds with a scalar, affine function G. In particular, the condition (V3) implies that $\Phi : K \to \mathbb{R}$ is continuous.

In the following chapters, we will mainly concentrate on the case where Φ is piecewise quadratic,

(V3)' $\Phi(z) = \frac{1}{2}b_i z^2 - f_i z + c_i, \quad \theta_i \le z \le \theta_{i+1}, \quad i = 0, \ldots, N,$

on a partition

$$\inf K = \theta_0 < \theta_1 < \ldots < \theta_N < \theta_{N+1} = \sup K$$

of the interval K. It is easily seen that (V3)' implies (V3), but, for example, $\Phi(z) = z^{1+\frac{1}{m}}$, $m \ge 1$, satisfies (V3) on $K = [0, +\infty)$ and is not quadratic.

Observe that the obstacle problem (1.8) is a (very simple) special case of our minimization problem (2.33). Indeed, after a suitable transformation of u, (1.8) can be rewritten in the form (2.33) with $\Phi \equiv 0$ on $K = (-\infty, 0]$. The bilinear form $a(\cdot,\cdot)$ defined in (1.6) is clearly elliptic on $H = H_0^1(\Omega)$ and the corresponding linear functional ℓ is bounded for sufficiently regular f (e.g. $f \in L^2(\Omega)$).

We will now state some properties of the functional ϕ as resulting from the properties (V1)–(V3) of Φ. First, let us recollect some standard notation from convex analysis. The functional ϕ defined on H is *convex* if

$$\phi(\omega v + (1-\omega)v') \le \omega\phi(v) + (1-\omega)\phi(v'), \quad \forall \omega \in [0,1], \quad \forall v, v' \in H.$$

A subset $\mathcal{K} \subset H$ is convex if the *indicator functional* $\chi_{\mathcal{K}}$, given by $\chi_{\mathcal{K}}(v) = 0$, $\forall v \in \mathcal{K}$, and $\chi_{\mathcal{K}}(v) = \infty$, $\forall v \notin \mathcal{K}$, is convex. ϕ is called *lower semicontinuous* if the convergence $v_k \to v$, $k \to \infty$, in H implies $\liminf_{k\to\infty} \phi(v_k) \ge \phi(v)$. We say that ϕ is *proper* if $\phi(v) > -\infty$, $\forall v \in H$, and $\phi \not\equiv \infty$. Finally, the subset $\operatorname{dom}\phi = \{v \in H \mid \phi(v) < +\infty\} \subset H$ is the *domain* of ϕ. As usual, a sequence $(v_k)_{k\ge 0}$ is said to *converge weakly* to $v \in H$, i.e. $v_k \rightharpoonup v$, $k \to \infty$, if $a(v_k, w) \to a(v, w)$, $k \to \infty$, holds for all $w \in H$.

Proposition 1.1 *The functional $\phi : H \to \mathbb{R} \cup \{+\infty\}$ is convex, lower semicontinuous and proper.*

The domain of ϕ is given by the non–empty, closed, and convex set

$$\mathcal{K} = \{v \in H \mid v(x) \in K \ \textit{a.e. in } \Omega\} \tag{2.38}$$

and $\phi : \mathcal{K} \to \mathbb{R}$ is continuous. Moreover, we have

$$v_k \rightharpoonup v, k \to \infty \quad \Rightarrow \quad \liminf_{k\to\infty} \phi(v_k) \ge \phi(v) \tag{2.39}$$

for all $(v_k)_{k\ge 0} \subset H$ and $v \in H$.

Proof. Let us first investigate the set $\mathcal{K}$. $\mathcal{K}$ is non–empty, because $0 \in K$ (see (V2)) provides $0 \in \mathcal{K}$. $\mathcal{K}$ is convex, because K is an interval and therefore convex. We now show that $\mathcal{K}$ is weakly closed and thus closed. Without loss of generality, let $K = (-\infty, 0]$ and consider a sequence $(v_k)_{k\geq 0} \subset \mathcal{K}$ such that $v_k \rightharpoonup v$, $k \to \infty$. Using the compact embedding of H in $L^2(\Omega)$, we obtain $v_k \to v$, $k \to \infty$, in $L^2(\Omega)$. We assume that $v \notin \mathcal{K}$. Then we can find a subset Ω' with positive measure such that $v(x) > 0$, $x \in \Omega'$, giving $\|v\|_{L^2(\Omega')} > 0$. Using $v_k \in \mathcal{K}$, we get $v(x) - v_k(x) \geq v(x)$, a. e. in Ω'. This leads to

$$\|v - v_k\|_{L^2(\Omega)} \geq \|v - v_k\|_{L^2(\Omega')} \geq \|v\|_{L^2(\Omega')} > 0, \quad \forall k \geq 0,$$

in contradiction to $v_k \to v$ in $L^2(\Omega)$.

In the next step, we show $\mathcal{K} = \operatorname{dom} \phi$. Let us first state that

$$\|G(v)\|_{L^2(\Omega)} \leq c\|v\|_{L^2(\Omega)} + C, \quad \forall v \in L^2(\Omega), \tag{2.40}$$

holds with $G(v)(x) = G(v(x))$, $x \in \Omega$, where G is the affine function from (V3). The constants c, C depend only on G and the (finite) measure of Ω. Now let $v \in \mathcal{K}$. Then $\phi(v) \in \mathbb{R}$ follows from

$$\begin{aligned} |\phi(v)| &\leq \textstyle\int_\Omega |\Phi(v(x))| dx \leq \int_\Omega (|\Phi(0)| + G(|v(x)|)|v(x)|) dx \\ &\leq |\Omega||\Phi(0)| + \|G(|v|)\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} < \infty, \end{aligned}$$

where (V3), the Cauchy–Schwarz inequality and (2.40) have been applied. It is clear that $\phi(v) = \infty$ if $v \notin \mathcal{K}$ so that $\mathcal{K} = \operatorname{dom} \phi$.

In order to demonstrate that ϕ is continuous on $\mathcal{K}$, we establish the stronger result

$$(v_k)_{k\geq 0} \subset \mathcal{K}, \;\; v_k \rightharpoonup v, \; k \to \infty \;\; \Rightarrow \;\; \phi(v_k) \to \phi(v), \; k \to \infty. \tag{2.41}$$

Let $(v_k)_{k\geq 0} \subset \mathcal{K}$ and $v_k \rightharpoonup v$, $k \to \infty$. We already know that $\mathcal{K}$ is weakly closed so that $v \in \mathcal{K}$. Again, we get the convergence $v_k \to v$, $k \to \infty$, in

$L^2(\Omega)$ from the compact embedding of H. Now (V3), the Cauchy–Schwarz inequality and (2.40) yield

$$\begin{aligned} |\phi(v_k) - \phi(v)| &\le \int_\Omega |\Phi(v_k(x)) - \Phi(v(x))|\, dx \\ &\le \int_\Omega G(|v_k(x)| + |v(x)|)|v_k(x) - v(x)|\, dx \\ &\le \|G(|v_k| + |v|)\|_{L^2(\Omega)} \|v_k - v\|_{L^2(\Omega)} \to 0, \end{aligned}$$

since the norms $\|v_k\|_{L^2(\Omega)}$, $k \ge 0$, are uniformly bounded.

We now prove (2.39). Let $v_k \rightharpoonup v$, $k \to \infty$. Assume that for each $k_0 \in \mathbb{N}$, there is an index $k \ge k_0$ such that $v_k \in \mathcal{K}$. Then we can find a subsequence $(v_{k_i})_{i\ge 0} \subset \mathcal{K}$ still converging weakly to v. Hence, $v \in \mathcal{K}$ and (2.41) yields $\phi(v_k) \to \phi(v)$, $k \to \infty$. In the remaining case, $v_k \notin \mathcal{K}$, $\forall k \ge k_0$, holds with some fixed $k_0 \ge 0$. Then we clearly have $\liminf_{k\to\infty} \phi(v_k) = \infty \ge \phi(v)$.

Finally, it follows from (V1), (2.39) and $\phi(v) \in \mathbb{R} \cup \{+\infty\}$, $\forall v \in H$, together with $\mathcal{K} = \operatorname{dom} \phi \ne \emptyset$ that ϕ is convex, lower semicontinuous and proper. This concludes the proof. □

We will only need ϕ be convex, lower semicontinuous and proper in order to ensure existence and uniqueness of the solution of our minimization problem (2.33). The additional results stated in Proposition 1.1 will be useful for the analysis of the finite element discretization later on.

1.2.2 Variational Inequalities

Throughout this section, we only assume that H is a Hilbert space with scalar product $a(\cdot,\cdot)$ and that the functional $\phi : H \to \mathbb{R} \cup \{\infty\}$ is convex, lower semicontinuous, and proper. In particular, all results to be derived in the sequel can be directly applied to the finite element discretization of the minimization problem (2.33).

The *epigraph* $\operatorname{epi} \phi$ of ϕ is defined by

$$\operatorname{epi} \phi = \{(v,s) \in H \times \mathbb{R} \mid \phi(v) \le s\} \subset H \times \mathbb{R}. \tag{2.42}$$

On the product space $H \times \mathbb{R}$, we introduce the canonical scalar product

$$(\boldsymbol{v}, \boldsymbol{w}) = a(v,w) + st, \quad \boldsymbol{v} = (v,s),\ \boldsymbol{w} = (w,t) \in H \times \mathbb{R},$$

and the corresponding norm $|\!|\!| \boldsymbol{v} |\!|\!| = (\boldsymbol{v}, \boldsymbol{v})^{1/2}$.

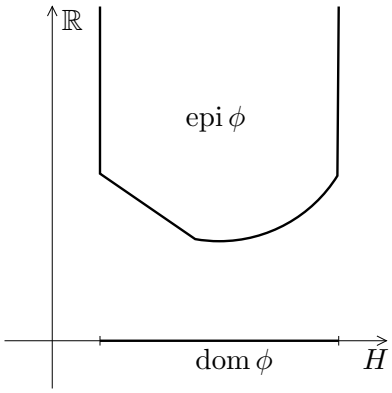

Figure 1.6 The epigraph epi ϕ

Lemma 1.2 *The epigraph* epi ϕ *is convex and closed in* $H \times \mathbb{R}$.

Proof. To see that epi ϕ is convex, choose $\omega \in [0,1]$ and (v,s), $(w,t) \in \operatorname{epi}\phi$. Then

$$\omega(v,s) + (1-\omega)(w,t) = (\omega v + (1-\omega)w, \omega s + (1-\omega)t) \in \operatorname{epi}\phi,$$

because $\phi(\omega v + (1-\omega)w) \le \omega s + (1-\omega)t$ follows from the convexity of ϕ. Consider a sequence $\boldsymbol{v}_k = (v_k, s_k) \in \operatorname{epi}\phi$, $k \ge 0$, converging to $\boldsymbol{v} = (v,s)$ in $H \times \mathbb{R}$. Then, we have $v_k \to v$ and $s_k \to s$, $k \to \infty$, so that we obtain

$$s = \lim_{k\to\infty} s_k \ge \liminf_{k\to\infty} \phi(v_k) \ge \phi(v),$$

because ϕ is lower semicontinuous. □

Vice versa, ϕ is convex and lower semicontinuous if epi ϕ is convex and closed in $H \times \mathbb{R}$ (see e.g. Aubin [4]).

Let us recall a well–known result on best approximation in Hilbert spaces.

Lemma 1.3 *Let* $\boldsymbol{v}_0 \in H \times \mathbb{R}$. *Then the minimization problem*

$$\boldsymbol{w} \in \operatorname{epi}\phi: \quad \| \boldsymbol{w} - \boldsymbol{v}_0 \| \le \| \boldsymbol{v} - \boldsymbol{v}_0 \|, \quad \forall\, \boldsymbol{v} \in \operatorname{epi}\phi, \tag{2.43}$$

has a unique solution $\boldsymbol{w}$. *Moreover,* $\boldsymbol{w}$ *satisfies the variational inequality*

$$(\, \boldsymbol{w} - \boldsymbol{v}_0, \boldsymbol{w} - \boldsymbol{v}\,) \le 0, \quad \forall\, \boldsymbol{v} \in \operatorname{epi}\phi. \tag{2.44}$$

Proof. In the first step, we show that (2.43) has a unique solution. Observe that $\operatorname{epi}\phi \neq \emptyset$, because ϕ is proper. Let $(\boldsymbol{w}_k)_{k\geq 0} \subset \operatorname{epi}\phi$ be a minimizing sequence, i.e.

$$\| \boldsymbol{w}_k - \boldsymbol{v}_0 \| \to \inf_{v\in\operatorname{epi}\phi} \| \boldsymbol{v} - \boldsymbol{v}_0 \| = \gamma \geq 0, \quad k \to \infty.$$

Inserting $\boldsymbol{v} = \boldsymbol{v}_0 - \boldsymbol{w}_i$ and $\boldsymbol{w} = \boldsymbol{v}_0 - \boldsymbol{w}_k$ in the so–called median formula

$$\| \boldsymbol{v} + \boldsymbol{w} \|^2 + \| \boldsymbol{v} - \boldsymbol{w} \|^2 = 2 \| \boldsymbol{v} \|^2 + 2 \| \boldsymbol{w} \|^2,$$

and using the convexity of $\operatorname{epi}\phi$, we get

$$\| \boldsymbol{w}_i - \boldsymbol{w}_k \|^2 \leq 2 \| \boldsymbol{w}_i - \boldsymbol{v}_0 \|^2 + 2 \| \boldsymbol{w}_k - \boldsymbol{v}_0 \|^2 - 4\gamma^2.$$

As the right–hand side tends to zero, $(\boldsymbol{w}_k)_{k\geq 0}$ is a Cauchy sequence and therefore convergent in $H \times \mathbb{R}$. The limit $\boldsymbol{w}^* \in H \times \mathbb{R}$ is contained in $\operatorname{epi}\phi$, because $\operatorname{epi}\phi$ is closed. Hence, we have

$$\gamma \leq \| \boldsymbol{w}^* - \boldsymbol{v}_0 \| \leq \| \boldsymbol{w}^* - \boldsymbol{w}_k \| + \| \boldsymbol{w}_k - \boldsymbol{v}_0 \| \to \gamma, \quad k \to \infty,$$

so that $\boldsymbol{w} = \boldsymbol{w}^*$ is a solution of (2.43). It is straightforward to see that this solution is unique.

Let $\omega \in (0, 1]$ and $\boldsymbol{v} \in \operatorname{epi}\phi$. Then $\boldsymbol{w} + \omega(\boldsymbol{v} - \boldsymbol{w}) \in \operatorname{epi}\phi$ and (2.43) yields

$$\begin{aligned} \| \boldsymbol{w} - \boldsymbol{v}_0 \|^2 &\leq \| \boldsymbol{w} + \omega(\boldsymbol{v} - \boldsymbol{w}) - \boldsymbol{v}_0 \|^2 \\ &= \| \boldsymbol{w} - \boldsymbol{v}_0 \|^2 + 2\omega \, (\boldsymbol{w} - \boldsymbol{v}_0, \boldsymbol{v} - \boldsymbol{w}) + \omega^2 \| \boldsymbol{v} - \boldsymbol{w} \|^2 \end{aligned}$$

so that we get

$$(\boldsymbol{w} - \boldsymbol{v}_0, \boldsymbol{w} - \boldsymbol{v}) - \tfrac{\omega}{2} \| \boldsymbol{w} - \boldsymbol{v} \|^2 \leq 0.$$

Inserting $\omega = 0$, we obtain (2.44). This completes the proof. □

Note that the variational inequality (2.44) is even equivalent to the minimization problem (2.43). We will present a more general result later on.

The following proposition is a key result of this section.

Proposition 1.4 *For each $\boldsymbol{v}_0 \in H \times \mathbb{R}$, with $\boldsymbol{v}_0 \notin \operatorname{epi}\phi$, there is an element $\boldsymbol{w}_0 \in H \times \mathbb{R}$ and $\varepsilon > 0$, such that*

$$(\boldsymbol{w}_0, \boldsymbol{v}) \leq (\boldsymbol{w}_0, \boldsymbol{v}_0) - \varepsilon, \qquad \forall\, \boldsymbol{v} \in \operatorname{epi}\phi. \tag{2.45}$$

Proof. Let $\boldsymbol{v}_0 \in H \times \mathbb{R}$, with $\boldsymbol{v}_0 \notin \operatorname{epi}\phi$. According to Lemma 1.3, we can find $\boldsymbol{w} \in \operatorname{epi}\phi$ satisfying the variational inequality (2.44). Denoting $\boldsymbol{w}_0 = \boldsymbol{v}_0 - \boldsymbol{w}$, we can rewrite (2.44) as

$$(\boldsymbol{w}_0, \boldsymbol{v}) \leq (\boldsymbol{w}_0, \boldsymbol{w}) = (\boldsymbol{w}_0, \boldsymbol{v}_0) - \|\boldsymbol{w} - \boldsymbol{v}_0\|^2, \quad \forall\, \boldsymbol{v} \in \operatorname{epi}\phi.$$

Hence, the assertion follows with $\varepsilon = \|\boldsymbol{w} - \boldsymbol{v}_0\|^2 > 0$. □

There is also a nice geometrical interpretation of Proposition 1.4. For each point $\boldsymbol{v}_0$ in the complement of $\operatorname{epi}\phi$, we can find a hyperplane

$$G = \{\boldsymbol{v} \in H \times \mathbb{R} \mid (\boldsymbol{w}_0, \boldsymbol{v}) = c\}$$

which *separates* $\boldsymbol{v}_0$ from $\operatorname{epi}\phi$ as illustrated in Figure 1.7.

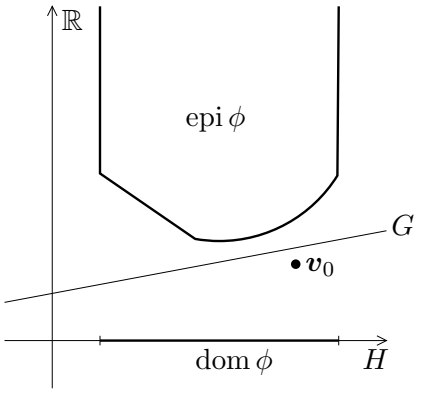

Figure 1.7 Separation of $\operatorname{epi}\phi$ and $\boldsymbol{v}_0$

The first separation theorems are due to Minkowski. The generalization of these theorems to Banach spaces gave rise to the problem of extending continuous linear forms which was finally settled by the celebrated Hahn–Banach theorem.

We now state an important consequence of Proposition 1.4, namely that the functional ϕ has an affine minorant. As we will see later, this provides a uniform lower bound for $\mathcal{J} + \phi$.

Proposition 1.5 *There are constants α_0, β_0 such that*

$$\phi(v) \geq \alpha_0 - \beta_0 \|v\|, \qquad \forall v \in H. \tag{2.46}$$

Proof. Let $v_0 \in \operatorname{dom}\phi$ and $s_0 < \phi(v_0)$ so that $\boldsymbol{v}_0 = (v_0, s_0) \notin \operatorname{epi}\phi$. According to Proposition 1.4, we can find $\boldsymbol{w}_0 = (w_0, t_0) \in H \times \mathbb{R}$ and $\varepsilon > 0$ such that

$$(\, \boldsymbol{w}_0, \boldsymbol{v} \,) \leq (\, \boldsymbol{w}_0, \boldsymbol{v}_0 \,) - \varepsilon, \qquad \forall\, \boldsymbol{v} \in \operatorname{epi}\phi. \tag{2.47}$$

In the first step, we show that $t_0 < 0$. Inserting $\boldsymbol{v} = (v_0, \phi(v) + s) \in \operatorname{epi}\phi$, $s \geq 0$, in (2.47), we obtain

$$st_0 \leq (\phi(v_0) + s_0)t_0 - \varepsilon < \infty.$$

If $t_0 > 0$, we can let $s \to \infty$ to get a contradiction, and $t_0 = 0$ would imply $\varepsilon \leq 0$. Hence, we have shown $t_0 < 0$.

Of course, (2.46) is trivial for $v \notin \operatorname{dom}\phi$. For arbitrary $v \in \operatorname{dom}\phi$, we insert $\boldsymbol{v} = (v, \phi(v)) \in \operatorname{epi}\phi$ in (2.47) and divide by t_0 to obtain

$$\phi(v) \geq t_0^{-1}\,(a(w_0, v_0) + s_0 t_0 - \varepsilon) - t_0^{-1} a(w_0, v).$$

Now the assertion follows from the Cauchy–Schwarz inequality. □

We are ready to state the main result of this section

Theorem 1.6 *The minimization problem (2.33) has a unique solution u and is equivalent to the variational inequality*

$$u \in H: \quad a(u, v-u) + \phi(v) - \phi(u) \geq \ell(v-u), \quad \forall v \in H. \tag{2.48}$$

Proof. First, we show the equivalence of (2.33) and the variational inequality (2.48). Let $u \in H$ be a solution of the minimization problem (2.33). Then, for arbitrary $v \in H$, we insert $u + \omega(v-u)$ with $\omega \in (0, 1]$ in (2.33) and use the convexity of ϕ to obtain

$$\begin{aligned}0 &\le \omega^{-1}(\mathcal{J}(u+\omega(v-u)) + \phi(u+\omega(v-u)) - \mathcal{J}(u) - \phi(u))\\ &\le \omega^{-1}(\tfrac{1}{2}\|u+\omega(v-u)\|^2 - \tfrac{1}{2}\|u\|^2) + \phi(v) - \phi(u) - \ell(v-u)\\ &= a(u, v-u) + \phi(v) - \phi(u) - \ell(v-u) + \tfrac{1}{2}\omega\|v-u\|^2.\end{aligned}$$

Now (2.48) follows as $\omega \to 0$.

Let $u \in H$ be a solution of the variational inequality (2.48). Using the estimate

$$0 \le \tfrac{1}{2}a(v-u, v-u) = \tfrac{1}{2}a(v,v) - a(u,v) + \tfrac{1}{2}a(u,u),$$

the inequality (2.48) then leads to

$$\begin{aligned}&\mathcal{J}(v) + \phi(v) - (\mathcal{J}(u) + \phi(u))\\ &\qquad = \tfrac{1}{2}a(v,v) - \tfrac{1}{2}a(u,u) + \phi(v) - \phi(u) - \ell(v-u)\\ &\qquad \ge a(u, v-u) + \phi(v) - \phi(u) - \ell(v-u) \ge 0, \quad \forall v \in H.\end{aligned}$$

Hence, u is a solution of (2.33).

In order to prove existence and uniqueness, we start by showing that $\mathcal{J} + \phi$ has a uniform lower bound. This follows immediately from the estimate (2.46) in Proposition 1.5, giving

$$\mathcal{J}(v) + \phi(v) \ge \tfrac{1}{2}\|v\|^2 - \|\ell\|\|v\| + \alpha_0 - \beta_0\|v\|, \quad \forall v \in H.$$

As in the proof of Lemma 1.3, we now show that a minimizing sequence $(v_k)_{k\ge 0}$ of $\mathcal{J} + \phi$, i.e. a sequence with the property

$$\mathcal{J}(v_k) + \phi(v_k) \to \inf_{v\in H}(\mathcal{J}(v) + \phi(v)) = \gamma > \infty, \quad k \to \infty,$$

converges to a solution of (2.33). Using the median formula

$$\|v_i - v_k\|^2 = 2\|v_i\|^2 + 2\|v_k\|^2 - 4\|\frac{v_i + v_k}{2}\|^2,$$

it follows from straightforward computation that

$$\begin{aligned}\tfrac{1}{4}\|v_i - v_k\|^2 &= \mathcal{J}(v_i) + \phi(v_i) - \gamma + \mathcal{J}(v_k) + \phi(v_k) - \gamma \\ &\quad +2\left(\gamma - \mathcal{J}(\tfrac{v_i+v_k}{2}) + \phi(\tfrac{v_i+v_k}{2})\right) \\ &\quad +2\left(\phi(\tfrac{v_i+v_k}{2}) - \tfrac{1}{2}(\phi(v_i) + \phi(v_k))\right) \\ &\le \mathcal{J}(v_i) + \phi(v_i) - \gamma + \mathcal{J}(v_k) + \phi(v_k) - \gamma,\end{aligned}$$

because γ is the infimum of $\mathcal{J}+\phi$ and ϕ is convex. By construction, the right–hand side tends to zero as $i,k \to \infty$. Hence, $(v_k)_{k\ge 0}$ is a Cauchy sequence and therefore convergent to some $u^* \in H$. To show that $u = u^* \in H$ is a solution of (2.33), observe that the lower semicontinuity of ϕ yields

$$\gamma \le \mathcal{J}(u^*) + \phi(u^*) \le \liminf_{k\to\infty}(\mathcal{J}(v_k) + \phi(v_k)) = \gamma.$$

Assume that there is another solution u'. Then u and u' must satisfy the variational inequality (2.48). Inserting $v = u'$ in the inequality for u and vice versa, we can sum up the resulting two estimates to obtain

$$0 \le a(u, u' - u) + a(u', u - u') = -\|u - u'\|^2.$$

This completes the proof. □

The variational inequality (2.48) is a generalization of the linear variational problem $a(u, v) = \ell(v)$, $\forall v \in H$. The inequality is the price we have to pay for circumventing the differentiation of the non–smooth functional ϕ. A different way of obtaining a variational formulation of the minimization problem (2.33) will be presented in the following section.

1.2.3 Subdifferentials

A bounded linear functional g on H with the property

$$\phi(v) - \phi(v_0) \ge \langle g, v - v_0\rangle, \qquad \forall v \in H, \tag{2.49}$$

is called a *subgradient* of ϕ at $v_0 \in H$. The *subdifferential* $\partial\phi(v_0)$ at v_0 is the set of all subgradients of ϕ at v_0. As a consequence, $\partial\phi$ can be regarded

as a *multivalued* function, or briefly a *multifunction*, which is defined on $\operatorname{dom}\partial\phi = \{v \in H \mid \partial\phi(v) \neq \emptyset\} \subset H$ and takes values in the set of subsets of the dual space H^*. It is easy to see that $\operatorname{dom}\partial\phi \subset \operatorname{dom}\phi$.

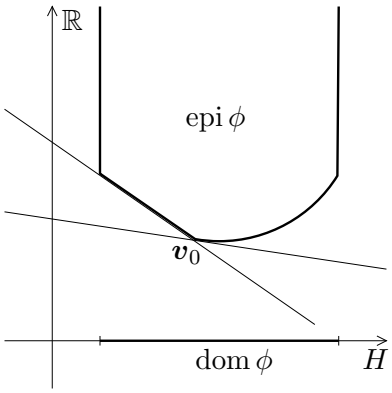

Figure 1.8 Supporting hyperplanes of $\operatorname{epi}\phi$ at $\boldsymbol{v}_0$

Observe that each subgradient $g \in \partial\phi(v_0)$ defines a *supporting hyperplane*

$$G = \{(v,t) \in H \times \mathbb{R} \mid -\langle g, v\rangle + t = -\langle g, v_0\rangle + \phi(v_0)\}$$

to the epigraph of ϕ at $\boldsymbol{v}_0 = (v_0, \phi(v_0))$ (see Figure 1.8). If ϕ is differentiable at $\boldsymbol{v}_0$, then the only supporting hyperplane at $\boldsymbol{v}_0$ is spanned by $(\phi'(v_0), 1)$. Hence, we have $\partial\phi(v_0) = \{\phi'(v_0)\}$ in this case, illustrating that the subdifferential is an extension of the usual derivative.

It follows immediately from the definition (2.49) that the multifunction $\partial\phi$ is *monotone* in the sense that

$$\langle g - g', v - v'\rangle \geq 0, \quad \forall g \in \partial\phi(v), \ \ \forall g' \in \partial\phi(v'),$$

holds for all $v_0, v_0' \in \operatorname{dom}\partial\phi$. As ϕ is lower semicontinuous, there is no monotone extension of $\partial\phi$, i.e.

$$\langle g - g', v - v'\rangle \geq 0, \quad \forall g \in \partial\phi(v), \ \ \forall v \in \operatorname{dom}\partial\phi,$$

implies $v' \in \operatorname{dom}\partial\phi$ and $g' \in \partial\phi(v')$ (see e.g. Deimling [43]). Such multifunctions are called *maximal monotone*.

An immediate consequence of the above considerations is the following

Proposition 1.7 *The minimization problem (2.33) is equivalent to the variational inclusion*

$$u \in H: \quad 0 \in a(u,v) - \ell(v) + \partial\phi(u)(v), \qquad \forall v \in H, \tag{2.50}$$

which has a unique solution $u \in H$.

Proof. As $\mathcal{J}$ is differentiable on H, it is easily checked that

$$\partial(\mathcal{J}+\phi)(v_0)(v) = a(v_0,v) - \ell(v) + \partial\phi(v_0)(v), \quad \forall v \in H, \tag{2.51}$$

holds for all $v_0 \in \operatorname{dom}\partial\phi$. If $u \in H$ solves the minimization problem (2.33), then we clearly have

$$\mathcal{J}(v) + \phi(v) - (\mathcal{J}(u) + \phi(u)) \geq 0, \qquad \forall v \in H,$$

so that $0 \in \partial(\mathcal{J}+\phi)(u)$ by definition. The other direction follows in a similar way. Finally, we know from Theorem 1.6 that (2.33) is uniquely solvable so that the same holds true for the variational inclusion (2.50). □

Observe that (2.50) contains the linear variational problem $a(u,v) = \ell(v)$, $\forall v \in H$, as a special case.

Proposition 1.7 motivates a further investigation of the subdifferential $\partial\phi$. For this reason, let us define the subdifferential $\partial\Phi$ of the scalar function Φ in the same way as above. Then it is straightforward to see that

$$(\partial\Phi(v_0), \cdot)_{L^2(\Omega)} \subset \partial\phi(v_0), \tag{2.52}$$

where the notation $\partial\Phi(v_0) = \{w \in L^2(\Omega) \mid w(x) \in \partial\Phi(v_0(x)) \text{ a.e. on } \Omega\}$ should not lead to confusion with the scalar multifunction $\partial\Phi$. As a consequence of (2.52), each solution of

$$u \in H: \quad 0 \in a(u,v) - \ell(v) + (\partial\Phi(u), v)_{L^2(\Omega)}, \quad \forall v \in H, \tag{2.53}$$

is a solution of the minimization problem (2.33), while the converse is not immediately clear. The following existence result was stated by Jerome [78], Proposition 3.2.1, p. 93, in a much more general form (see also Barbu [15] or Brézis [34, 35]).

Proposition 1.8 *The variational inclusion (2.53) has a solution.*

As a consequence of Propositions 1.7 and 1.8, the semi–discrete Stefan problem (1.22) has a unique solution which can be equivalently computed from the minimization problem (2.33) or from the variational inequality (2.48). In this case, the scalar function Φ is a primitive of the enthalpy $\mathcal{H}$, i.e. $\partial\Phi = \mathcal{H}$, having the properties (V1), (V2), and (V3)'. The same holds true for the semi–discrete porous medium equation (1.28), because Φ can be chosen such that $\partial\Phi = \mathcal{P}$ and (V1), (V2), and (V3) are valid.

Let us collect some further properties of scalar subdifferentials. Due to the conditions (V1)–(V3), we have

$$\operatorname{dom}\partial\Phi = \operatorname{dom}\Phi = K.$$

In fact, functions like $\Phi(z) = z^{1/2}$, $z \geq 0$, are excluded by (V3). Recall that the subdifferential ∂F of a convex, lower semicontinuous, and proper scalar function $F : \mathbb{R} \to \mathbb{R} \cup +\infty$ is maximal monotone. Conversely, each scalar maximal monotone multifunction is a subdifferential (see e.g. Barbu [15], p. 60). In general, the subdifferential of a sum of functions is not the sum of subdifferentials (see e.g. Deimling [43], p. 282). But assuming that F_1, $F_2 : \mathbb{R} \to \mathbb{R} \cup +\infty$ are convex, lower semicontinuous, and proper functions, which are *continuous* on their domain, we get

$$\partial F_1 + \partial F_2 = \partial(F_1 + F_2). \tag{2.54}$$

The following location principle will be useful later on.

Lemma 1.9 *Assume that $F : \mathbb{R} \to \mathbb{R} \cup +\infty$ is convex, lower semicontinuous, and proper. Let z_0, $z_1 \in \mathbb{R}$ such that*

$$z_0 + \inf\partial F(z_0) \leq 0 \quad \textit{if } z_0 \in \operatorname{dom}\partial F, \qquad z_0 \leq \inf\operatorname{dom}\partial F \quad \textit{else},$$
$$0 \leq z_1 + \sup\partial F(z_1) \quad \textit{if } z_1 \in \operatorname{dom}\partial F, \qquad z_1 \geq \sup\operatorname{dom}\partial F \quad \textit{else}.$$

Then there is a unique $\xi \in \operatorname{dom}\partial F$, such that $z_0 \leq \xi \leq z_1$ and $0 \in \xi + \partial F(\xi)$.

Proof. Using scalar versions of Theorem 1.6 and Proposition 1.7, the existence of a solution $\xi \in \operatorname{dom}\partial F$ with $0 \in \xi + \partial F(\xi)$ is immediately clear.

If z_0, $z_1 \in \operatorname{dom}\partial F$, then $z_0 \leq \xi \leq z_1$ follows from the monotonicity of ∂F.

If $\operatorname{dom}\partial F$ is bounded from above, i.e. if we have $\sup\operatorname{dom}\partial F = \overline{z} < \infty$, then $\lim_{z\to\overline{z}} \sup\partial F(z) = \infty$. Otherwise, there would be a monotone extension of ∂F. Similarly, we get $\lim_{z\to\underline{z}} \inf\partial F(z) = -\infty$, if $\inf\operatorname{dom}\partial F = \underline{z} > -\infty$. Using these observations, the remaining cases can be treated as above. □

1.3 Finite Element Discretization

We present a finite element discretization of the continuous minimization problem (2.33) introduced in Section 1.2.1. The solution space H is replaced by the discrete space of piecewise linear finite elements and the functional ϕ is approximated by a quadrature formula in order to separate the unknowns. The existence and uniqueness of discrete solutions follows from the general results in Section 1.2.2. Adapting the techniques presented by Glowinski [61] to the actual situation, we prove convergence in H. Error estimates are available for various special cases. For an extensive overview of the obstacle problem, we refer to Ciarlet [39], Section 23. Finite elements of higher order are also considered there. Error estimates for the semi–discrete Stefan problem are given by Elliott [50].

1.3.1 The Discrete Problem

Let $\mathcal{T}_j$ be a given partition of Ω in triangles $t \in \mathcal{T}_j$ with minimal diameter of order $\mathcal{O}(2^{-j})$. The sets of interior vertices and edges of $\mathcal{T}_j$ are called $\mathcal{N}_j$ and $\mathcal{E}_j$, respectively. We assume that each triangulation $\mathcal{T}_j$ is regular in the sense that the intersection of two triangles t, $t' \in \mathcal{T}_j$ consists of a common edge, a common vertex or is empty. A forbidden situation is illustrated in Figure 1.9

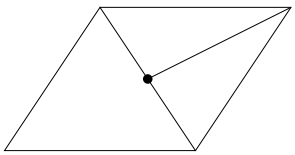

Figure 1.9 Forbidden irregular vertex

The finite element space $\mathcal{S}_j \subset H$ consists of all continuous functions $v \in H$ which are linear on each triangle $t \in \mathcal{T}_j$. $\mathcal{S}_j$ is spanned by the nodal basis

$$\Lambda_j = \{\lambda_p^{(j)} \mid p \in \mathcal{N}_j\}.$$

whose elements $\lambda_p^{(j)} \in \mathcal{S}_j$ are characterized by $\lambda_p^{(j)}(q) = \delta_{p,q}$, $\forall p, q \in \mathcal{N}_j$, (Kronecker delta).

We approximate the integral occurring in the definition of the functional ϕ in (2.37) by the quadrature formula resulting from the $\mathcal{S}_j$–interpolation of the integrand $\Phi(v)$ to obtain the *discrete functional* $\phi_j : \mathcal{S}_j \to \mathbb{R} \cup \{+\infty\}$,

$$\phi_j(v) = \sum_{p \in \mathcal{N}_j} \Phi(v(p))\, h_p, \tag{3.55}$$

with weights h_p defined by

$$h_p = \int_\Omega \lambda_p^{(j)}(x)\, dx.$$

Observe that the domain of ϕ_j is given by the non–empty, closed and convex set $\mathcal{K}_j \subset \mathcal{S}_j$,

$$\mathcal{K}_j = \{v \in \mathcal{S}_j \mid v(p) \in K,\ \forall p \in \mathcal{N}_j\}, \tag{3.56}$$

and that ϕ_j is continuous on $\mathcal{K}_j$.

Replacing the infinite–dimensional space H by $\mathcal{S}_j$ and the functional ϕ by ϕ_j, we end up with the *discrete minimization problem*

$$u_j \in \mathcal{S}_j : \quad \mathcal{J}(u_j) + \phi_j(u_j) \le \mathcal{J}(v) + \phi_j(v), \quad \forall v \in \mathcal{S}_j. \tag{3.57}$$

Let us consider the existence and uniqueness of a discrete solution u_j of (3.57). It is easily seen that the discrete functional ϕ_j is still convex, lower semicontinuous and proper. In the light of Section 1.2.2, we get the following discrete analogue of Theorem 1.6.

Theorem 1.10 *The discrete minimization problem (3.57) is equivalent to the discrete variational inequality*

$$u_j \in \mathcal{S}_j :\ a(u_j, v - u_j) + \phi_j(v) - \phi_j(u_j) \ge \ell(v - u_j),\ \forall v \in \mathcal{S}_j \tag{3.58}$$

and has a unique solution $u_j \in \mathcal{S}_j$.

As in the continuous case, we can rewrite (3.58) as a *discrete variational inclusion*

$$u_j \in \mathcal{S}_j : \quad 0 \in a(u_j, v) - \ell(v) + \partial\phi_j(u_j)(v), \quad \forall v \in \mathcal{S}_j. \tag{3.59}$$

However, we now have $\operatorname{dom} \partial\phi_j = \operatorname{dom} \phi_j$ together with the representation formula

$$\partial\phi_j(v_0)(v) = \sum_{p \in \mathcal{N}_j} \partial\Phi(v_0(p))v(p)\, h_p, \qquad \forall v \in \mathcal{S}_j, \tag{3.60}$$

which holds for all $v_0 \in \operatorname{dom} \partial\phi_j$.

1.3.2 Convergence Results

Let $\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_j$ be a sequence of triangulations with *decreasing mesh size*

$$h_j = \max_{t \in \mathcal{T}_j} \operatorname{diam} t \ \to \ 0, \quad j \to \infty.$$

In addition, we assume that the sequence $(\mathcal{T}_j)_{j \geq 0}$ is *shape regular* in the sense that the minimal interior angle of all occurring triangles is uniformly bounded from below.

The *consistency* of the discrete functionals ϕ_j, $j \geq 0$, is the subject of the next lemma. A variant of the density result is given by Glowinski [61], p. 36.

Lemma 1.11 *The subset $C_0^\infty(\Omega) \cap \mathcal{K}$ is dense in $\mathcal{K}$.*

Let $v \in C_0^\infty(\Omega)$ and define $v_j = I_{\mathcal{S}_j} v \in \mathcal{S}_j$ by piecewise linear interpolation. Then

$$v_j \to v \quad \textit{and} \quad \phi_j(v_j) \to \phi(v), \quad j \to \infty. \tag{3.61}$$

Proof. We first show that $C_0^\infty(\Omega) \cap \mathcal{K}$ is dense in $\mathcal{K}$. Let $v \in \mathcal{K} \subset H$. By definition of $H = H_0^1(\Omega)$, v is the limit of a sequence $(v_k)_{k \geq 0} \subset C_0^\infty(\Omega)$, but it is not self–evident that we can also ensure $v_k \in \mathcal{K}$, $\forall k \geq 0$. The construction of such a sequence can be carried out as follows. Let $\varphi \in C_0^\infty(\mathbb{R}^2)$ be a

mollifier, i.e. $\varphi(x) \geq 0$, $\int_{\mathbb{R}^2} \varphi(x)dx = 1$, and supp $\varphi \subset \{x \in \mathbb{R}^2 \mid |x| < 1\}$. Then we extend v by zero to $v \in H^1(\mathbb{R}^2)$, choose $\varepsilon > 0$, and define

$$w_\varepsilon(x) = \varepsilon^{-2} \int_{\mathbb{R}^2} v(\xi)\varphi(\frac{x-\xi}{\varepsilon})\, d\xi, \qquad \forall x \in \mathbb{R}^2.$$

It is well–known (cf. e.g. Adams [1], p. 29) that $w_\varepsilon \in C_0^\infty(\mathbb{R}^2)$ and $w_\varepsilon \to v$ as $\varepsilon \to 0$. Moreover, it is straightforward to see that

$$\inf K \leq \inf_{\xi\in\Omega} v(\xi) \leq w_\varepsilon(x) \leq \sup_{\xi\in\Omega} v(\xi) \leq \sup K, \quad \forall x \in \mathbb{R}^2,$$

and supp $w_\varepsilon \subset \Omega_\varepsilon = \{x \in \mathbb{R}^2 \mid \text{dist}(x, \Omega) < \varepsilon\}$.

In the next step, we introduce $v_\varepsilon \in C_0^\infty(\Omega) \cap \mathcal{K}$ by

$$v_\varepsilon(x) = w_\varepsilon(T_\varepsilon(x)), \qquad \forall x \in \Omega,$$

using an infinitely derivable transformation $T_\varepsilon : \Omega \to \mathbb{R}^2$ with $\Omega_\varepsilon \subset T_\varepsilon(\Omega)$, $T_\varepsilon(x) = x$ if $\text{dist}(x, \partial\Omega) \geq \varepsilon$, and

$$0 < c \leq |T_\varepsilon'(x)| \leq C < \infty, \quad \forall x \in \Omega, \;\; \varepsilon > 0.$$

Such a transformation can be easily constructed in the one–dimensional case and is then extended to $\Omega \subset \mathbb{R}^2$, taking into account that Ω has a polygonal boundary. It follows from the properties of T_ε that $w_\varepsilon - v_\varepsilon \to 0$ as $\varepsilon \to 0$, giving $\overline{C_0^\infty(\Omega) \cap \mathcal{K}} = \mathcal{K}$.

From now on, let $v \in C_0^\infty(\Omega)$ and $v_j = I_{\mathcal{S}_j} v$, $j \geq 0$. It is well–known that $v_j \to v$, $j \to \infty$ (see e.g. Ciarlet [39]). Note that the shape regularity of $(\mathcal{T}_j)_{j\geq 0}$ comes in here.

We show that $\phi_j(v_j) \to \phi(v)$, $j \to \infty$. Assume first that $\phi(v) = \infty$. Then we can find an open subset $\Omega' \subset \Omega$ with $v(x) \notin K$, $x \in \Omega'$. As $h_j \to 0$, it is clear that $\mathcal{N}_j \cap \Omega' \neq \emptyset$ holds for all $j \geq j_0$ with a suitable $j_0 \geq 0$. Hence, $\phi_j(v_j) \to \infty$, $j \to \infty$.

In the remaining case, we have $\phi(v) < \infty$ or, equivalently, $v \in \mathcal{K}$. As v has compact support in Ω with values in K and Φ is continuous on K, the

composition $\Phi(v(\cdot))$ is uniformly continuous on Ω. Let $p_1, p_2, p_3 \in \mathcal{N}_j$ denote the vertices of $t \in \mathcal{T}_j$. Then

$$\begin{aligned} |\phi(v) - \phi_j(v_j)| &\le \sum_{t\in\mathcal{T}_j} \int_t \left(\sum_{i=1}^{3} \lambda_{p_i}^{(j)}(x) |\Phi(v(x)) - \Phi(v(p_i))| \right) dx \\ &\le |\Omega| \max_{\{x,\xi\in\Omega,\, |x-\xi|\le h_j\}} |\Phi(v(x)) - \Phi(v(\xi))| \to 0, \quad j \to \infty. \end{aligned}$$

□

In addition to Lemma 1.11, we will need the following *stability* results.

Lemma 1.12 *There are constants α_0, β_0 such that*

$$\phi_j(v_j) \ge \alpha_0 - \beta_0 \|v_j\|, \quad \forall v_j \in S_j, \quad \forall j \ge 0. \tag{3.62}$$

Let $v_j \in \mathcal{S}_j$, $\forall j \ge 0$, and $v \in H$. Then

$$v_j \rightharpoonup v, \;\; j \to \infty \quad \Rightarrow \quad \liminf_{j\to\infty} \phi_j(v_j) \ge \phi(v). \tag{3.63}$$

Proof. The convexity of the scalar function Φ implies

$$\phi_j(v_j) \ge \phi(v_j), \quad \forall v_j \in \mathcal{S}_j, \quad \forall j \ge 0. \tag{3.64}$$

Now the assertions are an immediate consequence of Propositions 1.1 and 1.5. □

We are ready to state the main result of this section.

Theorem 1.13 *The solutions u_j of the discrete minimization problem (3.57) converge to the solution u of (2.33) in the sense that*

$$u_j \to u \quad \textit{and} \quad \phi_j(u_j) \to \phi(u), \qquad j \to \infty. \tag{3.65}$$

Proof. The proof is carried out in three steps.

First, we will show that $(u_j)_{j\geq 0}$ is bounded. Let $v \in C_0^\infty(\Omega) \cap \mathcal{K}$ and define $v_j = I_{\mathcal{S}_j} v \in \mathcal{S}_j$, $j \geq 0$, by interpolation. Since u_j satisfies the variational inequality (3.58), we have

$$\|u_j\|^2 = a(u_j, u_j) \leq a(u_j, v_j) + \phi_j(v_j) - \phi_j(u_j) - \ell(v_j - u_j).$$

We get uniform upper bounds for $\|v_j\|$ and $|\phi_j(v_j)|$ from Lemma 1.11 and have the uniform lower estimate (3.62) for $\phi_j(u_j)$. This leads to

$$\begin{aligned}\|u_j\|^2 &\leq a(u_j, v_j) + \phi_j(v_j) - \phi_j(u_j) - \ell(v_j - u_j)\\ &\leq \|u_j\|\|v_j\| + |\phi_j(v_j)| + |\alpha_0| + |\beta_0|\|u_j\| + \|\ell\|(\|v_j\| + \|u_j\|)\\ &\leq c\|u_j\| + C.\end{aligned}$$

As a consequence, $(u_j)_{j\geq 0}$ must be bounded.

In the next step, we show the weak convergence $u_j \rightharpoonup u$, $j \to \infty$, in H. As $(u_j)_{j\geq 0}$ is bounded, there is a subsequence $(u_{j_k})_{k\geq 0}$ and $u^* \in H$, such that $u_{j_k} \rightharpoonup u^*$, $k \to \infty$, in H. In order to prove $u^* = u$, we show that u^* is a solution of the variational inequality (2.48). Let $v \in C_0^\infty(\Omega)$ and $v_j = I_{\mathcal{S}_j} v \in \mathcal{S}_j$, $\forall j \geq 0$. Inserting v_j in the discrete variational inequality (3.58), we obtain

$$a(u_{j_k}, u_{j_k}) + \phi_{j_k}(u_{j_k}) \leq a(u_{j_k}, v_{j_k}) + \phi_{j_k}(v_{j_k}) - \ell(v_{j_k} - u_{j_k}).$$

The consistency (3.61) of ϕ_j and the weak convergence of u_{j_k} imply

$$\liminf_{k\to\infty} (a(u_{j_k}, u_{j_k}) + \phi_{j_k}(u_{j_k})) \leq a(u^*, v) + \phi(v) - \ell(v - u^*). \quad (3.66)$$

Utilizing

$$0 \leq a(u_{j_k} - u^*, u_{j_k} - u^*) = a(u^*, u^*) - 2a(u_{j_k}, u^*) + a(u_{j_k}, u_{j_k})$$

and the weak convergence of u_{j_k}, we deduce

$$a(u^*, u^*) \leq \liminf_{k\to\infty} a(u_{j_k}, u_{j_k}).$$

In connection with Lemma 1.12, this leads to

$$a(u^*, u^*) + \phi(u^*) \le \liminf_{k\to\infty} \left(a(u_{j_k}, u_{j_k}) + \phi_{j_k}(u_{j_k})\right). \tag{3.67}$$

Combining the estimates (3.66) and (3.67), we have shown

$$a(u^*, v - u^*) + \phi(v) - \phi(u^*) \ge \ell(v - u^*), \quad \forall v \in C_0^\infty(\Omega). \tag{3.68}$$

We will use a density argument to extend (3.68) to all $v \in H$. Assume that we can find a $v \in H$ such that (3.68) is wrong. Then we get $\phi(v) < \infty$, because $\phi(u^*) < \infty$ is clear from (3.68). Hence, $v \in \mathcal{K}$. According to Lemma 1.11, there is a sequence $(v_k)_{k\ge0} \subset C_0^\infty(\Omega) \cap \mathcal{K}$ converging to v. As ϕ is continuous on $\mathcal{K}$ (cf. Proposition 1.1), we also have $\phi(v_k) \to \phi(v)$, $k \to \infty$. Now the contradiction follows in the usual way. Hence, $u^* = u$ is the unique solution of the variational inequality (2.48) and we obtain $u_j \rightharpoonup u$, $j \to \infty$.

Finally, we will prove the strong convergence of $(u_j)_{j\ge0}$. Again, consider some fixed $v \in C_0^\infty(\Omega)$ and let $v_j = I_{\mathcal{S}_j} v \in \mathcal{S}_j$, $\forall j \ge 0$. Using the discrete variational inequality (3.58), we compute

$$\begin{aligned} \|u - u_j\|^2 + \phi_j(u_j) &\le a(u,u) - 2a(u,u_j) + a(u_j,u_j) + \phi_j(u_j) \\ &\le a(u,u) - 2a(u,u_j) + a(u_j,v_j) + \phi_j(v_j) - \ell(v_j - u_j). \end{aligned} \tag{3.69}$$

The right–hand side of (3.69) converges to $a(u, v-u) + \phi(v) - \ell(v-u)$ as $j \to \infty$. Hence, we obtain

$$\begin{aligned} \liminf_{j\to\infty} \phi_j(u_j) &\le \liminf_{j\to\infty} \left(\|u - u_j\|^2 + \phi_j(u_j)\right) \\ &\le \limsup_{j\to\infty} \left(\|u - u_j\|^2 + \phi_j(u_j)\right) \\ &\le a(u, v-u) + \phi(v) - \ell(v-u), \quad \forall v \in C_0^\infty(\Omega). \end{aligned} \tag{3.70}$$

We apply the same density argument as above to extend (3.70) to all $v \in H$. Inserting $v = u$ in (3.70) and using (3.63), we get

$$\phi(u) \le \liminf_{j\to\infty} \phi_j(u_j) \le \limsup_{j\to\infty} \left(\|u - u_j\|^2 + \phi_j(u_j)\right) \le \phi(u).$$

This provides the convergence results (3.65). □

Error estimates are known for various special cases. For obstacle problems, we get $\|u-u_j\| = \mathcal{O}(h_j)$, if the data are sufficiently regular. Even for smooth data, the regularity of the solution is limited by $u \in W^{s,p}(\Omega)$, $s < 2 + 1/p$, $1 < p < \infty$, so that piecewise quadratic finite elements only give $\mathcal{O}(h_j^{3/2-\varepsilon})$, $\varepsilon > 0$, (cf. Brezzi, Hager and Raviart [36]). Optimal error estimates for the spatial problems arising from a time-discretized Stefan problem were given by Elliott [50].

The following chapters will be devoted to the fast solution of the discrete problem (3.57) and to the construction of a suitable triangulation. It will turn out that both problems are closely related.

2 Relaxation Methods

In the previous chapter, we analyzed the convex minimization problem (1.2.33) and derived a convergent finite element approximation. Basic assumptions and notations are stated in Sections 1.2.1 and 1.3.1. Now we will focus on the iterative solution of the resulting discrete minimization problem (1.3.57). Throughout the remainder of this work, we assume that the scalar function Φ generating the functional ϕ_j is piecewise quadratic, i.e. we require the sharper condition (V3)' in Section 1.2.1, p. 23, instead of (V3). This is not essential for the basic convergence results to be presented in this chapter, but it will simplify the construction of monotone multigrid methods later on.

Relaxation methods of nonlinear Gauß–Seidel type have been well understood since the early eighties (see e.g. Glowinski [61]). In particular, it is well–known that such single grid relaxations are globally convergent for the class of problems under consideration.

In the framework of successive subspace correction methods (cf. Xu [122]), Gauß–Seidel relaxations are generated by the direct splitting of the underlying finite element space $\mathcal{S}_j$ in the one–dimensional subspaces spanned by the high–frequency nodal basis functions $\lambda_p^{(j)} \in \Lambda_j$. This explains their unsatisfactory convergence rates caused by a bad representation of the low–frequency contributions of the error. Fast solvers, such as multigrid methods, can be derived by extending the splitting induced by Λ_j by additional subspaces spanned by suitable functions with large support.

In the case of linear selfadjoint elliptic problems, this point of view led to a new type of convergence proofs for multigrid methods. For an introduction to this field, we refer to the basic surveys of Bramble [31], Xu [122] and Yserentant [126]. See also the monographs of Griebel [63] and Oswald [104] in this series.

Here, the above reasoning motivates the introduction of *extended relaxation methods* (cf. Kornhuber [82, 83]) for the iterative solution of (1.3.57). Extended relaxation methods can be regarded as special variants of *nonlinear*

successive subspace correction methods or multilevel projection schemes (cf. McCormick [95]).

It will turn out that only approximate versions can be implemented with optimal numerical complexity. The global convergence of extended relaxations is preserved by local *monotone approximations*, as introduced in Section 2.2. In the final section of this chapter, we will give sufficient criteria for the asymptotic invariance of the discrete phases and we will show that *quasioptimal* monotone approximations asymptotically provide the same convergence rates as the original scheme.

2.1 Basic Convergence Results

The basic idea of relaxation methods is to decompose the global minimization problem (1.3.57) in a number of local subproblems. The convergence speed of the resulting iterative scheme depends heavily on the choice of the underlying decomposition of $\mathcal{S}_j$. After a brief introduction to well–known convergence results on single grid relaxations of Gauß–Seidel type, we will introduce extended relaxation methods. Such schemes preserve the global convergence and on the other hand give much more flexibility in the choice of the decomposition of $\mathcal{S}_j$.

2.1.1 Gauß–Seidel Relaxation

The (nonlinear) *Gauß–Seidel relaxation method* results from the successive optimization of the energy functional $\mathcal{J} + \phi_j$ in the direction of $\lambda_p^{(j)} \in \Lambda_j$. Recall that

$$\Lambda_j = (\lambda_{p_1}^{(j)}, \ldots, \lambda_{p_{n_j}}^{(j)})$$

is the nodal basis of the finite element space $\mathcal{S}_j$. Observe that the nodal basis functions $\lambda_p^{(j)}$, $p \in \mathcal{N}_j$, are *high–frequency* functions.

To give a precise formulation, we introduce the splitting

$$\mathcal{S}_j = \sum_{l=1}^{n_j} V_l, \quad V_l = \operatorname{span}\{\lambda_{p_l}^{(j)}\}, \tag{1.1}$$

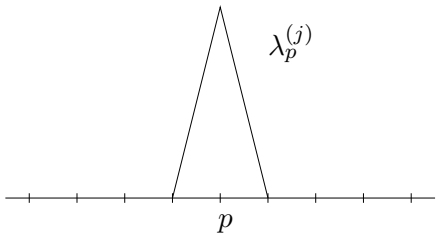

Figure 2.1 High–frequency nodal basis function $\lambda_P^{(j)} \in \Lambda_j$

of $\mathcal{S}_j$ in the one–dimensional subspaces $V_l \subset \mathcal{S}_j$. Then, starting with a given iterate $w_0^\nu = u_j^\nu \in \mathcal{S}_j$, $\nu \geq 0$, we compute a sequence of intermediate iterates w_l^ν from the n_j local subproblems

$$\begin{aligned} \bar{v}_l^\nu \in V_l: \quad & \mathcal{J}(w_{l-1}^\nu + \bar{v}_l^\nu) + \phi_j(w_{l-1}^\nu + \bar{v}_l^\nu) \\ & \leq \mathcal{J}(w_{l-1}^\nu + v) + \phi_j(w_{l-1}^\nu + v), \quad \forall v \in V_l, \end{aligned} \tag{1.2}$$

setting $w_l^\nu = w_{l-1}^\nu + \bar{v}_l^\nu$, $l = 1, \ldots, n_j$. Finally, the next iterate $u_j^{\nu+1}$ is given by

$$u_j^{\nu+1} = \mathcal{M}_j(u_j^\nu) = w_{n_j}^\nu = u_j^\nu + \sum_{l=1}^{n_j} \bar{v}_l^\nu. \tag{1.3}$$

Note that the possible acceleration of the convergence speed by multiplying the corrections with additional relaxation factors is not considered here. For notational convenience, the index ν will frequently be skipped in the sequel.

In the light of Theorem 1.6, each of the local subproblems (1.2) is *uniquely solvable* and can be equivalently rewritten as the variational inequality

$$\begin{aligned} \bar{v}_l \in V_l: \quad & a(\bar{v}_l, v - \bar{v}_l) + \phi_j(w_{l-1} + v) - \phi_j(w_{l-1} + \bar{v}_l) \\ & \geq \ell(v - \bar{v}_l) - a(w_{l-1}, v - \bar{v}_l), \quad \forall v \in V_l. \end{aligned} \tag{1.4}$$

By construction, we have *monotonically decreasing energy*,

$$\mathcal{J}(w_l) + \phi_j(w_l) \leq \mathcal{J}(w_{l-1}) + \phi_j(w_{l-1}), \quad l = 1, \ldots, n_j, \tag{1.5}$$

and the uniqueness of the correction $\bar{v}_l$ implies that equality holds if and only if $w_l = w_{l-1}$. This leads to

$$\mathcal{J}(\mathcal{M}_j(w)) + \phi_j(\mathcal{M}_j(w)) = \mathcal{J}(w) + \phi_j(w) \quad \Leftrightarrow \quad \mathcal{M}_j(w) = w. \tag{1.6}$$

Observe that for arbitrary initial iterate $u_j^0 \in \mathcal{S}_j$ the first iterate u_j^1 has *finite energy* or, equivalently,

$$\mathcal{M}_j(w) \in \mathcal{K}_j = \operatorname{dom} \phi_j, \quad \forall w \in \mathcal{S}_j. \tag{1.7}$$

We will see later on that the iteration operator $\mathcal{M}_j$ is continuous, i.e.

$$w_k \to w \quad \Rightarrow \quad \mathcal{M}_j(w_k) \to \mathcal{M}_j(w), \quad k \to \infty, \tag{1.8}$$

holds for every convergent sequence $(w_k)_{k\geq 0} \subset \mathcal{S}_j$.

Now we are ready to prove the global convergence of the Gauß–Seidel relaxation.

Theorem 2.1 *For any initial iterate $u_j^0 \in \mathcal{S}_j$, the sequence of iterates $(u_j^\nu)_{\nu\geq 0}$ provided by the Gauß–Seidel relaxation method (1.3) converges to the solution u_j of the discrete problem (1.3.57).*

Proof. The proof is divided into three steps. In the beginning, we show the sequence of iterates $(u_j^\nu)_{\nu\geq 0}$ is bounded. As ϕ_j has an affine minorant (cf. Lemma 1.12), we have

$$\mathcal{J}(v) + \phi_j(v) \geq \tfrac{1}{2}\|v\|^2 - c\|v\| - C, \quad \forall v \in \mathcal{S}_j, \tag{1.9}$$

so that $v \to \infty$ implies $\mathcal{J}(v)+\phi_j(v) \to \infty$. Hence, $(u_j^\nu)_{\nu\geq 0}$ must be bounded, because

$$\mathcal{J}(u_j^\nu) + \phi_j(u_j^\nu) \leq \mathcal{J}(u_j^1) + \phi_j(u_j^1) < \infty, \quad \forall \nu \geq 1.$$

Let $(u_j^{\nu_k})_{k\geq 0}$ be an arbitrary, convergent subsequence of $(u_j^\nu)_{\nu\geq 0}$,

$$u_j^{\nu_k} \to u^* \in \mathcal{S}_j, \quad k \to \infty.$$

Such a subsequence exists, because $(u_j^\nu)_{\nu\geq 0}$ is bounded and $\mathcal{S}_j$ has finite dimension. Moreover, we have $u^* \in \mathcal{K}_j$, because $u_j^{\nu_k} \in \mathcal{K}_j$, $\forall k \geq 1$, and $\mathcal{K}_j$

is a closed subset of $\mathcal{S}_j$. We now prove that u^* must be a fixed point of $\mathcal{M}_j$. For notational convenience, we use the abbreviation $\bar{\mathcal{J}} = \mathcal{J} + \phi_j$. The monotonicity (1.5) implies

$$\bar{\mathcal{J}}(u_j^{\nu_k+1}) = \bar{\mathcal{J}}(\mathcal{M}_j(u_j^{\nu_k})) \leq \bar{\mathcal{J}}(u_j^{\nu_k}), \quad \forall k \geq 0.$$

From (1.5), we also have

$$\bar{\mathcal{J}}(u_j^{\nu_{k+1}}) \leq \bar{\mathcal{J}}(u_j^{\nu_k+1}), \quad \forall k \geq 0.$$

By virtue of the continuity of $\mathcal{M}_j$ and the continuity of $\bar{\mathcal{J}}$ on $\mathcal{K}_j$, this leads to

$$\bar{\mathcal{J}}(\mathcal{M}_j(u^*)) = \bar{\mathcal{J}}(u^*), \tag{1.10}$$

and we conclude from (1.6) that $\mathcal{M}_j(u^*) = u^*$.

In the final step, we show that u_j is the only fixed point of $\mathcal{M}_j$. Let $\mathcal{M}_j(u^*) = u^*$. Then it is sufficient to prove

$$a(u^*, v_j - u^*) + \phi_j(v_j) - \phi_j(u^*) \geq \ell(v_j - u^*), \quad \forall v_j \in \mathcal{S}_j, \tag{1.11}$$

because we know from Theorem 1.10 that u_j is the unique solution of this variational inequality. Exploiting the special structure of the functional ϕ_j,

$$\phi_j(v) = \sum_{l=1}^{n_j} \Phi(v(p_l))\ h_{p_l}, \quad \forall v \in \mathcal{S}_j,$$

each of the local variational inequalities (1.4) takes the form

$$\begin{aligned} a(\bar{v}_l, v - \bar{v}_l) + \Phi(w_{l-1}(p_l) + v(p_l))h_{p_l} - \Phi(w_{l-1}(p_l) + \bar{v}_l(p_l))h_{p_l} \\ \geq \ell(v - \bar{v}_l) - a(w_{l-1}, v - \bar{v}_l), \quad \forall v \in V_l. \end{aligned}$$

As u^* is a fixed point of $\mathcal{M}_j$, all local corrections $\bar{v}_l$ of u^* must be zero so that

$$a(u^*, v) + \Phi(u^*(p_l) + v(p_l))h_{p_l} - \Phi(u^*(p_l))h_{p_l} \geq \ell(v) \tag{1.12}$$

holds for all $v \in V_l$ and $l = 1, \ldots, n_j$. Now consider some arbitrary but fixed $v_j \in \mathcal{S}_j$. Inserting the interpolation

$$v = I_{V_l}(v_j - u^*) = (v_j(p_l) - u^*(p_l))\lambda_{p_l}^{(j)} \in V_l$$

in (1.12), we obtain

$$a(u^*, I_{V_l}(v_j - u^*)) + \Phi(v_j(p_l))h_{p_l} - \Phi(u^*(p_l))h_{p_l} \geq \ell(I_{V_l}(v_j - u^*))$$

for $l = 1, \ldots, n_j$. Adding up all these local inequalities we get (1.11). This proves $u^* = u_j$.

We have shown that each convergent subsequence of $(u_j^\nu)_{\nu \geq 0}$ converges to u_j. Hence, the whole sequence must converge to u_j. This completes the proof. □

Observe that the proof makes strong use of the fact that the unknowns are decoupled with respect to ϕ_j. In fact, there are simple counterexamples (see e.g. Glowinski [61]) showing that this decoupling is necessary for the global convergence of relaxation methods of Gauß–Seidel type.

Up till now, we have not made use of condition (V3)' from Section 1.2.1, p. 23, stating that the functional ϕ_j is piecewise quadratic. Now we will exploit this property in order to derive an explicit formula for the corrections $\bar{v}_l$. In the light of Proposition 1.7, each of the local minimization problems (1.2) is equivalent to the variational inclusion

$$\begin{aligned} \bar{v}_l \in V_l: \quad 0 \in a(\bar{v}_l, v) - (\ell(v) - a(w_{l-1}, v)) \\ + \partial\phi_j(w_{l-1} + \bar{v}_l)(v), \quad \forall v \in V_l. \end{aligned} \tag{1.13}$$

It is clear that $\bar{v}_l \in V_l = \operatorname{span}\{\lambda_{p_l}^{(j)}\}$ can be written as

$$\bar{v}_l = \bar{z}_l \lambda_{p_l}^{(j)}$$

introducing the unknown *correction factor* $\bar{z}_l \in \mathbb{R}$. Hence, (1.13) can be reformulated as the scalar inclusion

$$\bar{z}_l \in \mathbb{R}: \quad 0 \in a_{ll}\bar{z}_l - r_l + \partial\Phi_l(\bar{z}_l), \tag{1.14}$$

where we have used the definitions

$$a_{ll} = a(\lambda_{p_l}^{(j)}, \lambda_{p_l}^{(j)}), \quad r_l = \ell(\lambda_{p_l}^{(j)}) - a(w_{l-1}, \lambda_{p_l}^{(j)}) \tag{1.15}$$

and the multifunction $\partial\Phi_l$ is the subdifferential of the scalar convex function

$$\Phi_l(z) = \phi_j(w_{l-1} + z\lambda_{p_l}^{(j)}), \qquad \forall z \in \mathbb{R}. \tag{1.16}$$

Observe that the subdifferential $\partial\Phi_l$ is given by

$$\partial\Phi_l(z) = \partial\Phi(w_{l-1}(p_l) + z)\; h_{p_l}, \quad \forall z \in \operatorname{dom}\partial\Phi_l. \tag{1.17}$$

As Φ is piecewise quadratic according to condition (V3)', p. 23, we have

$$\Phi(z) = \frac{1}{2}b_i z^2 - f_i z + c_i, \quad \theta_i \le z \le \theta_{i+1}, \quad i = 0, \ldots, N,$$

on a partition

$$\inf K = \theta_0 < \theta_1 < \ldots < \theta_N < \theta_{N+1} = \sup K$$

of the interval K (cf. condition (V2), p. 23). Hence, the subdifferential $\partial\Phi$ takes the form

$$\partial\Phi(z) = \begin{cases} b_i z - f_i & \text{if } \theta_i < z < \theta_{i+1},\ i = 0, \ldots, N \\ [s_i^-, s_i^+] & \text{if } \quad z = \theta_i, \quad i = 1, \ldots, N \end{cases}, \tag{1.18}$$

where we used the abbreviations

$$s_i^- = b_{i-1}\theta_i - f_{i-1} \le b_i\theta_i - f_i = s_i^+, \;\; i = 1, \ldots, N.$$

Note that the interval $[s_i^-, s_i^+]$ represents the jump of the derivative Φ' at the transition point θ_i. We introduce the partition

$$\vartheta_0^- \le \vartheta_0^+ \le \ldots \le \vartheta_i^- \le \vartheta_i^+ \le \ldots \le \vartheta_{N+1}^- \le \vartheta_{N+1}^+ \tag{1.19}$$

of the real axis $\mathbb{R}$. With the exception of $\vartheta_0^- = -\infty$ and $\vartheta_{N+1}^+ = +\infty$, the grid points ϑ_i^-, ϑ_i^+ are given by

$$\vartheta_i^- = a_{p_l}\theta_i + s_i^-, \;\; \vartheta_i^+ = a_{p_l}\theta_i + s_i^+, \quad i = 0, \ldots, N+1,$$

where we have set $a_{p_l} = a_{ll}/h_{p_l}$. Once we have determined an interval of the partition (1.19) containing the modified residual $r_{p_l} = (r_l + a_{ll}w_{l-1}(p_l))/h_{p_l}$, the solution $\bar{z}_l$ of (1.14) is obtained from

$$\bar{z}_l = -w_{l-1}(p_l) + \begin{cases} \theta_i, & \vartheta_i^- \le r_{p_l} \le \vartheta_i^+ \\ (r_{p_l} + f_i)/(a_{p_l} + b_i), & \vartheta_i^+ \le r_{p_l} \le \vartheta_{i+1}^- \end{cases}. \tag{1.20}$$

Observe that the right–hand side of (1.20) is a continuous function of w_{l-1}. This proves the continuity (1.8) of $\mathcal{M}_j$.

2.1.2 Extended Relaxation Methods

Though the Gauß–Seidel relaxation method is globally convergent, it usually provides unsatisfactory convergence rates for decreasing meshsize. To improve the speed of convergence, we now *extend* the set Λ_j by additional search directions.

Let $(M^\nu)_{\nu\geq 0}$ be a given sequence of ordered subsets $M^\nu = (\mu_1^\nu, \ldots, \mu_{m^\nu}^\nu)$ of $\mathcal{S}_j$, $\forall \nu \geq 0$. We assume that the leading elements of M^ν are the nodal basis functions,

$$\text{(M1)} \qquad M^\nu = (\lambda_{p_1}^{(j)}, \ldots, \lambda_{p_{n_j}}^{(j)}, \mu_{n_j+1}^\nu, \ldots, \mu_{m^\nu}^\nu), \quad \forall \nu \geq 0.$$

The elements of the extension $M_c^\nu = (\mu_{n_j+1}^\nu, \ldots, \mu_{m^\nu}^\nu)$ are intended to play the role of *coarse grid functions* with large support, in contrast to the *fine grid functions* contained in Λ_j. Note that the case $\mu_l^\nu = \mu_{l'}^\nu$, $l \neq l'$, is not excluded so that the same function may appear several times in each subset M^ν.

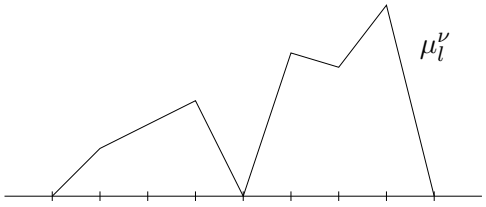

Figure 2.2 Low–frequency function $\mu_l^\nu \in M^\nu$

The *extended relaxation method induced by* $(M^\nu)_{\nu\geq 0}$ results from the successive minimization of the energy $\mathcal{J} + \phi_j$ in the direction of $\mu_l^\nu \in M^\nu$ for $\nu = 0, 1, \ldots$.

In order to describe one iteration step in detail, we assume that u_j^ν is given for some fixed $\nu \geq 0$. Then M^ν gives rise to the splitting

$$\mathcal{S}_j = \sum_{l=1}^{m^\nu} V_l^\nu, \quad V_l^\nu = \operatorname{span}\{\mu_l^\nu\}, \tag{1.21}$$

of $\mathcal{S}_j$ in one–dimensional subspaces $V_l^\nu \subset \mathcal{S}_j$. Starting with $w_0^\nu = u_j^\nu$, we compute a sequence of intermediate iterates w_l^ν from the m^ν local subproblems

$$\begin{aligned} \bar{v}_l^\nu \in V_l^\nu : \quad & \mathcal{J}(w_{l-1}^\nu + \bar{v}_l^\nu) + \phi_j(w_{l-1}^\nu + \bar{v}_l^\nu) \\ & \leq \mathcal{J}(w_{l-1}^\nu + v) + \phi_j(w_{l-1}^\nu + v), \ \ \forall v \in V_l^\nu, \end{aligned} \tag{1.22}$$

setting $w_l^\nu = w_{l-1}^\nu + \bar{v}_l^\nu$, $l = 1, \ldots, m^\nu$. Then the next iterate is given by

$$u_j^{\nu+1} = w_{m^\nu}^\nu = u_j^\nu + \sum_{l=1}^{m^\nu} \bar{v}_l^\nu. \tag{1.23}$$

To simplify the notation, the index ν will frequently be suppressed.

Extended relaxation methods can be equivalently characterized as a special type of *nonlinear* successive subspace correction method generated by *one–dimensional* splittings of the form (1.21). Observe that M_c^ν may change in each iteration step, so that (1.21) can be iteratively adapted to the discrete phases of the finite element approximation u_j.

The local corrections $\bar{v}_l$ in direction of $\lambda_{p_l}^{(j)} \in \Lambda_j$ and $\mu_l \in M_c$ are called *fine grid corrections* and *coarse grid corrections*, respectively. Observe that the leading fine grid corrections correspond to a Gauß–Seidel relaxation step. The resulting intermediate iterate is called *smoothed iterate* $\bar{u}_j^\nu = w_{n_j}^\nu$. We have seen above that $\mathcal{J}(\bar{u}_j^0) + \phi_j(\bar{u}_j^0) < \infty$, $\forall u_j^0 \in \mathcal{S}_j$. This leads to

$$0 \in \operatorname{dom} \phi_j(w_{l-1}^\nu + \cdot\,), \quad \forall l > n_j, \ \ \forall \nu \geq 0, \tag{1.24}$$

so that the translated functional $\phi_j(w_{l-1}^\nu + \cdot) : V_l \to \mathbb{R} \cup \{+\infty\}$ is convex, lower semicontinuous, and proper for all $l = 1, \ldots, m^\nu$ and all $\nu \geq 0$. Now it follows in the usual way that each of the local subproblems (1.22) is uniquely solvable.

By construction, the extended relaxation is locally monotone in the sense that

$$\mathcal{J}(w_l) + \phi_j(w_l) \leq \mathcal{J}(w_{l-1}) + \phi_j(w_{l-1}), \quad l = 1, \ldots, m^\nu. \tag{1.25}$$

We now introduce a damped version of (1.23). For a given iterate $w_0^\nu = u_j^\nu$, $\nu \geq 0$, the intermediate iterates $w_l^\nu = w_{l-1}^\nu + v_l^\nu$, $l = 1, \ldots, m^\nu$, of an

extended underrelaxation induced by $(M^\nu)_{\nu\geq 0}$ result from the exact fine grid corrections

$$v_l^\nu = \bar{v}_l^\nu, \qquad l = 1, \ldots, n_j, \tag{1.26}$$

and from the damped coarse grid corrections

$$v_l^\nu = \omega_l^\nu \bar{v}_l^\nu, \;\; \omega_l^\nu \in [0,1], \qquad l = n_j + 1, \ldots, m^\nu. \tag{1.27}$$

Each of the optimal corrections $\bar{v}_l^\nu$ is computed from (1.22). The next iterate $u_j^{\nu+1}$ of the extended underrelaxation is given by

$$u_j^{\nu+1} = w_{m^\nu}^\nu = u_j^\nu + \sum_{l=1}^{m^\nu} v_l^\nu. \tag{1.28}$$

Again, we will mostly skip the index ν.

The leading fine grid corrections are evaluated exactly. In particular, we still have $\operatorname{dom} \phi_j(w_{l-1} + \cdot) \neq \emptyset$ so that the local corrections v_l are well–defined.

Due to the convexity of $\mathcal{J} + \phi_j$, the monotonicity (1.25) is preserved by the damping (1.27). As a consequence, the global convergence is inherited from the nonlinear Gauß–Seidel scheme.

Theorem 2.2 *For any initial iterate* $u_j^0 \in \mathcal{S}_j$ *and any sequence of damping parameters* $(\omega_l^\nu)_{\nu\geq 0}$ *occurring in (1.27), the sequence of iterates* $(u_j^\nu)_{\nu\geq 0}$ *produced by the extended underrelaxation (1.28) converges to the solution* $u_j \in \mathcal{S}_j$ *of the finite element discretization (1.3.57).*

Proof. We proceed in the same way as in the convergence proof for the Gauß–Seidel relaxation. Again, we use the abbreviation $\bar{\mathcal{J}} = \mathcal{J} + \phi_j$. Exploiting

$$\bar{\mathcal{J}}(u_j^\nu) \leq \bar{\mathcal{J}}(\bar{u}_j^0) < \infty, \quad \forall \nu \geq 1,$$

together with (1.9), we conclude that $(u_j^\nu)_{\nu\geq 0}$ is bounded.

Hence, we can find a convergent subsequence $(u_j^{\nu_k})_{k\geq 0}$

$$u_j^{\nu_k} \to u^* \in \mathcal{K}_j, \quad k \to \infty.$$

Here, we used that $(u_j^{\nu_k})_{k\geq 1} \subset \mathcal{K}_j$ and that $\mathcal{K}_j$ is closed. In the next step, we show that u^* is a fixed point of $\mathcal{M}_j$. In fact, each step of an extended underrelaxation starts with a single grid relaxation, so that the local monotonicity (1.25) implies

$$\bar{\mathcal{J}}(u_j^{\nu_k+1}) \leq \bar{\mathcal{J}}(u_j^{\nu_k+1}) \leq \bar{\mathcal{J}}(\mathcal{M}_j(u_j^{\nu_k})) \leq \bar{\mathcal{J}}(u_j^{\nu_k}), \quad \forall k \geq 0,$$

and the continuity of $\mathcal{M}_j$ together with the continuity of $\bar{\mathcal{J}}$ on $\mathcal{K}_j$ yield

$$\bar{\mathcal{J}}(\mathcal{M}_j(u^*)) = \bar{\mathcal{J}}(u^*). \tag{1.29}$$

Using (1.6), we deduce $\mathcal{M}_j(u^*) = u^*$.

We already know that u_j is the only fixed point of $\mathcal{M}_j$ (cf. Theorem 2.1) so that $u^* = u_j$. As $(u_j^{\nu_k})_{k\geq 0}$ was an arbitrary convergent subsequence, the whole sequence of iterates must converge to u_j. This completes the proof. □

Later on, we will also need the convergence of the intermediate iterates w_l^ν.

Corollary 2.3 *For any initial iterate $u_j^0 \in \mathcal{S}_j$ and any sequence of damping parameters $(\omega_l^\nu)_{\nu\geq 0}$ occurring in (1.27), the sequence of intermediate iterates $(w_l^\nu)_{\nu\geq 0}$ produced by the extended underrelaxation (1.28) converges to the solution $u_j \in \mathcal{S}_j$ of the finite element discretization (1.3.57).*

Proof. Assume that there is a subsequence $w_k = w_{l_k}^{\nu_k}$, $\forall k \geq 0$, which is not converging to u_j. Then we first conclude as above that $(w_k)_{k\geq 0}$ is bounded. Therefore, we can find a further subsequence, still denoted by $(w_k)_{k\geq 0}$, which is converging to some $w^* \in \mathcal{S}_j$. We can additionally assume that

$$\bar{\mathcal{J}}(u_j^{\nu_k+1}) \leq \bar{\mathcal{J}}(w_k) \leq \bar{\mathcal{J}}(u_j^{\nu_k}), \quad \forall k \geq 0, \tag{1.30}$$

because the w_k are intermediate iterates. We clearly have $u_j^{\nu_k}, w_k \in \mathcal{K}_j, \forall k \geq 1$, and $\bar{\mathcal{J}}$ is continuous on $\mathcal{K}_j$. Hence, (1.30) implies that $\bar{\mathcal{J}}(w^*) = \bar{\mathcal{J}}(u_j)$. As u_j is the unique solution of the minimization problem (1.3.57), we get $w^* = u_j$. This completes the proof. □

Extended underrelaxations provide a general framework for the acceleration of convergent single grid relaxations by additional (coarse grid) corrections. *The essential point is that the additional corrections must not increase the energy.* As the convergence of nonlinear Gauß–Seidel relaxations is not restricted to the special situation considered here, Theorem 2.2 is open to

various generalizations. For example, only the explicit formula (1.20) for the fine grid correction factors requires that the scalar function Φ (generating ϕ_j by (1.3.64)) is piecewise quadratic. Moreover, Φ can be replaced by a family of different functions Φ_p, $\forall p \in \mathcal{N}_j$. Piecewise linear finite elements are also not essential. We will consider piecewise quadratic approximations in connection with a posteriori error estimates later on.

2.2 Monotone Approximations

While the leading fine grid corrections are given explicitly, the exact evaluation of the coarse grid corrections turns out to be too expensive in practical calculations. On the other hand, a suitable approximation of the local subproblems (1.22) should preserve the global convergence of the original extended relaxation method. In this section, we will introduce *monotone approximations* which have this property. The key observation is that monotone approximations provide an (implicit) damping of the exact coarse grid corrections.

Let $w_{l-1} \in \mathcal{S}_j$, $l > n_j$, be an intermediate iterate. Using the same arguments as in Section 2.1.1, we can compute the next coarse grid correction

$$\bar{v}_l = \bar{z}_l \mu_l \in V_l$$

from the scalar inclusion

$$\bar{z}_l \in \mathbb{R}: \quad 0 \in a_{ll}\bar{z}_l - r_l + \partial\Phi_l(\bar{z}_l). \tag{2.31}$$

Here, we have generalized the definition (1.15) of a_{ll} and r_l according to

$$a_{ll} = a(\mu_l, \mu_l), \quad r_l = \ell(\mu_l) - a(w_{l-1}, \mu_l), \tag{2.32}$$

and Φ_l is now defined by

$$\Phi_l(z) = \phi_j(w_{l-1} + z\mu_l), \qquad \forall z \in \mathbb{R}. \tag{2.33}$$

Recall that w_{l-1} has finite energy or, equivalently, $0 \in \operatorname{dom} \phi_j(w_{l-1} + \cdot)$ (see (1.7)). Hence, Φ_l is convex, lower semicontinuous, and proper so that the

subdifferential $\partial\Phi_l$ is maximal monotone. Using the representation (1.3.55) of ϕ_j, we obtain

$$\Phi_l(z) = \sum_{p\in\mathcal{N}_j} \Phi(w_{l-1}(p) + z\mu_l(p))\, h_p, \quad \forall z \in \mathbb{R}. \tag{2.34}$$

As Φ is continuous on $\operatorname{dom}\Phi = K$, the subdifferential of the sum can be taken elementwise (see (1.2.54)). Hence, we can write $\partial\Phi_l$ in the form

$$\partial\Phi_l(z) = \sum_{p\in\mathcal{N}_j} \mu_l(p)\, \partial\Phi(w_{l-1}(p) + z\mu_l(p))\, h_p, \ \ \forall z \in \operatorname{dom}\partial\Phi_l. \tag{2.35}$$

In principle, we can evaluate the solution of the scalar inclusion (2.31) in a similar way as in Section 2.1.1. However, the situation now is more complicated. The main reason is that the number of transition points of $\partial\Phi_l$ is no longer fixed, but grows with the number of nodes which are contained in the support of μ_l. Recall that $\operatorname{supp}\mu_l$ is assumed to be large for $\mu_l \in M_c$.

This motivates the approximation of the subdifferential $\partial\Phi_l^\nu$ occurring in (2.31) by a suitable maximal monotone multifunction $\partial\Psi_l^\nu$. Recall that a maximal monotone multifunction $\partial\Psi_l^\nu$ is always the subdifferential of a convex, lower semicontinuous, and proper functional $\Psi_l^\nu : \mathbb{R} \to \mathbb{R}\cup\{+\infty\}$. As usual, we will mostly skip the index ν.

Definition 2.4 *A scalar maximal monotone multifunction $\partial\Psi_l$ is called a* monotone approximation of $\partial\Phi_l$, *if*

$$0 \in \operatorname{dom}\partial\Psi_l \subset \operatorname{dom}\partial\Phi_l \tag{2.36}$$

and if the following estimates hold for all $z \in \operatorname{dom}\partial\Psi_l$,

$$\begin{aligned} &\mathit{sup}\ \partial\Psi_l(z) \ge \mathit{sup}\ \partial\Phi_l(z), \qquad z \ge 0, \\ &\mathit{inf}\ \partial\Psi_l(z) \le \mathit{inf}\ \partial\Phi_l(z), \qquad z \le 0. \end{aligned} \tag{2.37}$$

In particular, we have $\partial\Phi_l(0) \subset \partial\Psi_l(0)$. Note that the above conditions are trivially satisfied for $\partial\Psi_l = \partial\Psi_\infty$, given by $\partial\Psi_\infty(0) = \mathbb{R}$ on $\operatorname{dom}\partial\Psi_\infty = \{0\}$.

The approximation $\partial\Psi_l$ of $\partial\Phi_l$ gives rise to the approximate subproblem

$$z_l \in \mathbb{R}: \quad 0 \in a_{ll}z_l - r_l + \partial\Psi_l(z_l). \tag{2.38}$$

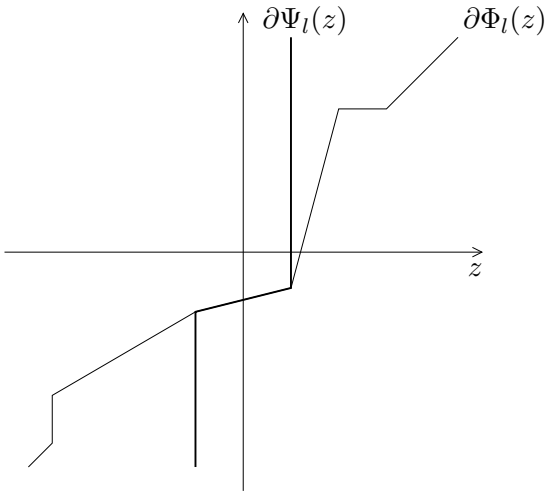

Figure 2.3 Monotone approximation $\partial\Psi_l$

It is clear from the assumptions on $\partial\Psi_l$ that (2.38) admits a unique solution $z_l \in \operatorname{dom} \partial\Psi_l$. The resulting approximate coarse grid correction is given by

$$v_l = z_l \mu_l \in V_l.$$

We are now going to clarify how the monotone approximation of $\partial\Phi_l$ affects the exact local corrections $\bar{v}_l$.

Proposition 2.5 *Assume that $\partial\Psi_l$ is a monotone approximation of $\partial\Phi_l$. Then the corrections $\bar{v}_l$ and v_l, computed from (2.31) and (2.38) respectively are related by*

$$v_l = \omega_l \bar{v}_l, \quad \omega_l \in [0,1]. \tag{2.39}$$

Proof. Assume that the solution $\bar{z}_l$ of (2.31) is non–negative. Then we set $z_0 = 0$, $z_1 = \bar{z}_l$, and $F(z) = (\Psi_l - r_l)/a_{ll}$ in order to apply Lemma 1.9. Using these definitions together with (2.36) and (2.37), we obtain $0 \le z_l = \xi \le \bar{z}_l$. The remaining case can be treated in a symmetrical way. □

Proposition 2.5 gives rise to the following definition.

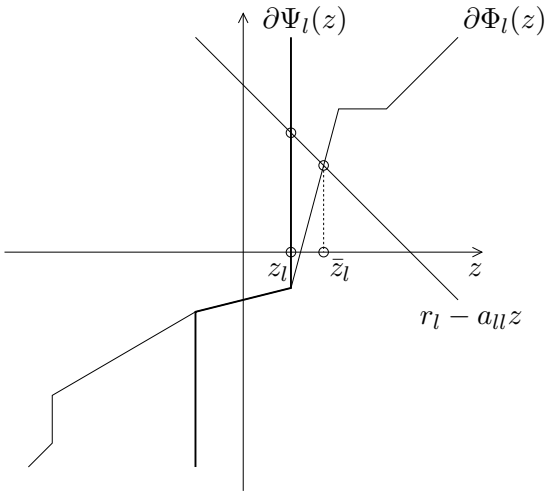

Figure 2.4 Implicit local damping

Definition 2.6 *Let $(\partial\Psi_l^\nu)_{\nu\geq 0}$ be a sequence of monotone approximations, where we have formally set*

$$\partial\Psi_l^\nu = \partial\Phi_l^\nu, \qquad l = 1, \ldots, n_j, \quad \forall \nu \geq 0.$$

Then, for given $u_j^0 \in \mathcal{S}_j$, the extended underrelaxation induced by $(M^\nu)_{\nu\geq 0}$ and the monotone approximations $(\partial\Psi_l^\nu)_{\nu\geq 0}$ *provides the iterates*

$$u_j^{\nu+1} = u_j^\nu + \sum_{l=1}^{m^\nu} v_l^\nu, \quad \forall \nu \geq 0,$$

where the local corrections v_l^ν are computed from the (approximate) local problems (2.38).

As a consequence of Proposition 2.5, we can apply Theorem 2.2 to show that the approximate iterative scheme introduced in Definition 2.6 is globally convergent. Corollary 2.3 implies that the sequence of intermediate iterates also converges to u_j. Note that the (local) relaxation parameters ω_l are only used in the analysis and do not appear in actual computations.

Though the original extended relaxation method and the approximate version are both globally convergent, the convergence rates of the approximate

version may be considerably worse. In particular, we have to exclude the trivial monotone approximation $\partial\Psi_\infty$, which brings back the simple Gauß–Seidel relaxation. In the next section, we will give sufficient criteria providing the same asymptotic convergence behavior of the monotone approximation and of the original scheme.

2.3 Asymptotic Properties: The Linear Reduced Problem

The *discrete phases* $\mathcal{N}_j^i(v) \subset \mathcal{N}_j$ of some fixed $v \in \mathcal{S}_j$ are given by

$$\mathcal{N}_j^i(v) = \{p \in \mathcal{N}_j \mid v(p) \in (\theta_i, \theta_{i+1})\}, \quad i = 0, \dots, N, \tag{3.40}$$

and we define $\mathcal{N}_j^\circ(v) = \bigcup_{i=1}^N \mathcal{N}_j^i(v)$. In the remaining *critical nodes* $\mathcal{N}_j^\bullet(v)$,

$$\mathcal{N}_j^\bullet(v) = \mathcal{N}_j \setminus \mathcal{N}_j^\circ(v), \tag{3.41}$$

the values of v are taken from the set $\{\theta_0, \theta_1, \dots, \theta_N, \theta_{N+1}\}$ of transition points.

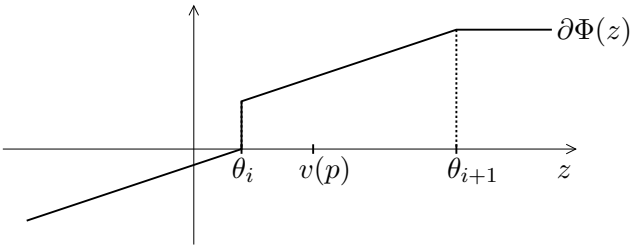

Figure 2.5 Discrete phases: $p \in \mathcal{N}_j^i(v)$

In the case of a discrete obstacle problem, $\mathcal{N}_j^\bullet(v)$ is equal to the discrete coincidence set v. For discrete versions of the semi–discrete Stefan problem (1.1.22), we have a discrete ice and a discrete water phase. The critical set $\mathcal{N}_j^\bullet(v)$ then consists of all nodes p where $v(p) = 0$.

In this section, we will first prove the convergence of the discrete phases of the (intermediate) iterates produced by extended underrelaxations. This will allow us to clarify further the asymptotic behavior of extended underrelaxations induced by monotone approximations (cf. Definitions 2.4, 2.6).

In advance, let us state some further conditions on the sequence $(M^\nu)_{\nu\geq 0}$ of search directions. We will require that $(M^\nu)_{\nu\geq 0}$ is positive and bounded in the sense that there are constants c, C, not depending on ν, such that

$$\text{(M2)}\quad 0 < c \leq \mu_l^\nu(p) \leq C, \quad \forall p \in \text{ int supp } \mu_l^\nu \cap \mathcal{N}_j, \ \ \forall \mu_l^\nu \in M^\nu, \ \ \forall \nu \geq 0.$$

Note that (M2) may be regarded as a stability condition on $(M^\nu)_{\nu\geq 0}$. Some sort of consistency is expressed by the condition that there is a sequence $M^* = (\mu_1^*, \ldots, \mu_{m^*}^*)$, not depending on ν, such that

$$\text{(M3)}\quad \mathcal{N}_j^\bullet(\bar{u}_j^\nu) = \mathcal{N}_j^\bullet(u_j) \ \ \Rightarrow \ \ M^\nu = M^*$$

holds for all $\nu \geq 0$. Recall that the smoothed iterate $\bar{u}_j^\nu$ results from the leading Gauß–Seidel relaxation step applied to u_j^ν.

The discrete problem (1.3.57) is called *non–degenerate*, if

$$p \in \mathcal{N}_j^\bullet(u_j) \Rightarrow \ell(\lambda_p^{(j)}) - a(u_j, \lambda_p^{(j)}) \in \text{ int } \partial\phi_j(u_j)(\lambda_p^{(j)}). \tag{3.42}$$

This condition describes the stability of the critical nodes $\mathcal{N}_j^\bullet(u_j)$ with respect to small perturbations of u_j. In the continuous case, related conditions are frequently used in the analysis of the regularity of the free boundary.

Proposition 2.7 *Assume that the discrete problem (1.3.57) is non–degenerate and that $(M^\nu)_{\nu\geq 0}$ satisfies the conditions (M1), cf. p. 51, and (M2). Then the phases of the intermediate iterates $(w_l^\nu)_{\nu\geq 0}$ resulting from the extended underrelaxation (1.28) induced by $(M^\nu)_{\nu\geq 0}$ converge to the phases of $u_j \in \mathcal{S}_j$ in the sense that*

$$\begin{aligned} &\mathcal{N}_j^i(w_l^\nu) = \mathcal{N}_j^i(u_j), \qquad \forall i = 0, \ldots, N, \\ &w_l^\nu(p) = u_j(p), \qquad\quad \forall p \in \mathcal{N}_j^\bullet(u_j), \end{aligned} \tag{3.43}$$

holds for all $l = 1, \ldots, m^\nu$ and all $\nu \geq \nu_0$ with a suitable $\nu_0 \geq 0$.

Proof. The convergence of the intermediate iterates $(w_l^\nu)_{\nu\geq 0}$ follows from Corollary 2.3. This implies that there is a $\nu_1 \geq 0$ with the property

$$\mathcal{N}_j^i(u_j) \subset \mathcal{N}_j^i(w_l^\nu), \quad \forall i = 0, \ldots, N, \ \ \forall \nu \geq \nu_1. \tag{3.44}$$

Using again the convergence of $(w_l^\nu)_{\nu\geq 0}$ and (3.44), it is sufficient to show that $\mathcal{N}_j^\bullet(u_j) \subset \mathcal{N}_j^\bullet(w_l^\nu)$ holds for large ν.

In the first step, we derive the extended non–degeneracy condition

$$\ell(\mu_l^\nu) - a(u_j, \mu_l^\nu) \in I_l \subset \ \text{int}\ \partial\phi_j(u_j)(\mu_l^\nu), \quad \forall \nu \geq 0, \tag{3.45}$$

for all $\mu_l^\nu \in M^\nu$ with the property $\ \text{int supp}\ \mu_l^\nu \cap \mathcal{N}_j^\bullet(u_j) \neq \emptyset$. The closed intervals $I_l \subset \mathbb{R}$ are defined by

$$I_l = \{z \in \mathbb{R} |\ |z - (\ell(\mu_l^\nu) - a(u_j, \mu_l^\nu))| \leq \varepsilon\}$$

and $\varepsilon > 0$ is independent of l or ν. Indeed, as a consequence of the non–degeneracy condition (3.42), we can find an $\varepsilon_j > 0$ such that (3.45) holds for all $\mu_l^\nu = \lambda_{p_l}^{(j)} \in \Lambda_j$. Taking the constant c from (M2), we can easily check that (3.45) is valid for all $\mu_l^\nu \in M^\nu$, if ε satisfies $0 < \varepsilon \leq c\, \varepsilon_j$.

Let us consider the fine grid correction in some fixed node $p_l \in \mathcal{N}_j^\bullet(u_j)$ with $u_j(p_l) = \theta_i$. Utilizing (3.45) and the convergence of $(w_l^\nu)_{\nu\geq 0}$, we can find a threshold $\nu_2 \geq \nu_1$ such that

$$\ell(\lambda_{p_l}^{(j)}) - a(w_l^\nu, \lambda_{p_l}^{(j)}) \in \ \text{int}\ \partial\phi_j(u_j)(\lambda_{p_l}^{(j)}), \quad \forall \nu \geq \nu_2, \tag{3.46}$$

holds for all $p_l \in \mathcal{N}_j^\bullet(u_j)$. Recall that w_l^ν results from the fine grid correction associated with $\lambda_{p_l}^{(j)}$. This property can be rewritten as

$$\ell(\lambda_{p_l}^{(j)}) - a(w_l^\nu, \lambda_{p_l}^{(j)}) \in \partial\phi_j(w_l^\nu)(\lambda_{p_l}^{(j)}). \tag{3.47}$$

Using the representation (1.2.49) of $\partial\phi_j$,

$$\partial\phi_j(w_{l-1}^\nu)(\lambda_{p_l}^{(j)}) = \partial\Phi(w_{l-1}^\nu(p_l))\ h_{p_l},$$

together with the monotonicity of $\partial\Phi$, we deduce from (3.46) and (3.47) that $w_l^\nu(p_l) = u_j(p_l) = \theta_i$.

Hence, the fine grid correction makes sure that, for large ν, each critical point of u_j is a critical point of the smoothed iterate $\bar{u}_j^\nu$. We still have to show that these critical points are not affected by the subsequent coarse grid correction, i.e. that

$$\text{int supp } \mu_l^\nu \cap \mathcal{N}_j^\bullet(u_j) \neq \emptyset \Rightarrow \bar{v}_l = 0, \quad \forall \mu_l^\nu \in M_c^\nu, \quad \forall \nu \geq \nu_3, \tag{3.48}$$

holds with a suitable $\nu_3 \geq \nu_2$. Let $\mu_l^\nu \in M_c^\nu$ and $\text{int supp } \mu_l^\nu \cap \mathcal{N}_j^\bullet(u_j) \neq \emptyset$. By virtue of (M1), we can assume inductively that the values of w_{l-1}^ν in $p \in \text{int supp } \mu_l^\nu \cap \mathcal{N}_j^\bullet(u_j)$ have been fixed by the leading fine grid corrections and have not been changed by subsequent coarse grid corrections. The convergence of $(w_l^\nu)_{\nu \geq 0}$ together with the upper bound in (M2) implies that

$$a(w_l^\nu, \mu_l^\nu) - a(u_j, \mu_l^\nu) \to 0, \quad \nu \to \infty.$$

Hence, we can use (3.45), the representation (1.2.49) of $\partial\phi_j$,

$$\partial(w_{l-1}^\nu)(\mu_l^\nu) = \sum_{p \in \mathcal{N}_j} \partial\Phi(w_{l-1}^\nu(p))\mu_l^\nu(p)\, h_p,$$

the upper bound in (M2), and the continuity of the derivative $\partial\Phi(z) = \Phi'(z)$ in $z \in \text{int } K \setminus \{\theta_1, \ldots, \theta_N\}$ to find a $\nu_3 \geq \nu_2$ such that

$$\ell(\mu_l^\nu) - a(w_{l-1}^\nu, \mu_l^\nu) \in \partial\phi_j(w_{l-1}^\nu)(\mu_l^\nu), \quad \forall \nu \geq \nu_3. \tag{3.49}$$

Using the "scalar" notation introduced in (2.32) and (2.33), the inclusion (3.49) can be rewritten as $r_l \in \partial\Phi_l(0)$, giving $\bar{z}_l = 0$. □

Once the phases $\mathcal{N}_j^i(u_j)$ of the exact solution u_j are known, we can define the bilinear form $b_{u_j}(v, w)$,

$$b_{u_j}(v, w) = \sum_{i=0}^{N} \sum_{p \in \mathcal{N}_j^i(u_j)} b_i v(p) w(p)\, h_p, \tag{3.50}$$

and the functional $f_{u_j}(v)$,

$$f_{u_j}(v) = \sum_{i=0}^{N} \sum_{p \in \mathcal{N}_j^i(u_j)} f_i v(p)\, h_p \tag{3.51}$$

on the finite element space $\mathcal{S}_j$. Denoting

$$a_{u_j}(v,w) = a(v,w) + b_{u_j}(v,w), \quad \ell_{u_j}(v) = \ell(v) + f_{u_j}(v), \tag{3.52}$$

we can easily check that $u_j = u_j^\circ$ is the solution of the reduced linear problem

$$u_j^\circ \in \bar{\mathcal{S}}_j^\circ : \quad a_{u_j}(u_j^\circ, v) = \ell_{u_j}(v), \quad \forall v \in \mathcal{S}_j^\circ, \tag{3.53}$$

where $\bar{\mathcal{S}}_j^\circ = \{v \in \mathcal{S}_j \mid v(p) = u_j(p),\ \forall p \in \mathcal{N}_j^\bullet(u_j)\}$ and the reduced subspace $\mathcal{S}_j^\circ \subset \mathcal{S}_j$ is defined by

$$\mathcal{S}_j^\circ = \{v \in \mathcal{S}_j \mid v(p) = 0,\ \forall p \in \mathcal{N}_j^\bullet(u_j)\}. \tag{3.54}$$

Consider $M^* = (\mu_1^*, \ldots, \mu_{m^*}^*)$ from (M3). Then the reduced set

$$M^\circ = M^* \cap \mathcal{S}_j^\circ,$$

induces a *linear* extended relaxation method

$$u_j^{\nu+1} = u_j^\nu + \sum_{l=1}^{m^*} \bar{v}_l \tag{3.55}$$

for the iterative solution of (3.53). The corrections $\bar{v}_l \in V_l$ in the direction of $\mu_l^* \in M^\circ$ are computed from the linear local subproblems

$$\bar{v}_l \in V_l : \quad a_{u_j}(\bar{v}_l, v) = \ell_{u_j}(v) - a(w_{l-1}, v), \quad \forall v \in V_l. \tag{3.56}$$

Assuming that the original discrete problem (1.3.57) is non–degenerate, one can show that an extended relaxation induced by a sequence $(M^\nu)_{\nu\geq 0}$ with the properties (M1)–(M3) reduces asymptotically to the linear scheme (3.55). In order to obtain a related result for extended underrelaxations induced by a sequence of monotone approximations $(\partial\Psi_l^\nu)_{\nu\geq 0}$, we have to impose further restrictions on the multifunctions $\partial\Psi_l^\nu$.

Definition 2.8 *The monotone approximations $(\partial\Psi_l^\nu)_{\nu\geq 0}$ of $(\partial\Phi_l^\nu)_{\nu\geq 0}$ are called* quasioptimal *if the convergence of the intermediate iterates $(w_l^\nu)_{\nu\geq 0}$ and of their phases implies that there is an index $\nu_0 \geq 0$ and an open interval $I \subset \mathbb{R}$, independent of ν and l, such that $0 \in I$ and*

$$\partial\Psi_l^\nu(z) = \partial\Phi_l^\nu(z), \quad \forall z \in I, \quad \forall \nu \geq \nu_0, \tag{3.57}$$

holds if $\mu_l^\nu(p) = 0$ for all $p \in \mathcal{N}_j^\bullet(u_j)$, $l = 1, \ldots, m^\nu$.

Now we are ready to state the main result of this chapter.

Theorem 2.9 *Assume that the discrete problem (1.3.57) is non–degenerate (cf. (3.42)) and that the sequence of search directions* $(M^\nu)_{\nu\geq 0}$ *satisfies the conditions (M1)–(M3). Then the extended underrelaxation induced by* $(M^\nu)_{\nu\geq 0}$ *and quasioptimal approximations* $(\partial\Psi_l^\nu)_{\nu\geq 0}$ *is globally convergent and is reducing to the linear extended relaxation (3.55) for* $\nu \geq \nu_0$ *with suitable* $\nu_0 \geq 0$.

Proof. The global convergence is clear from Theorem 2.2. It follows from Proposition 2.7 together with (M3) that $\mathcal{N}_j^\bullet(w_l^\nu) = \mathcal{N}_j^\bullet(u_j)$ and $M^\nu = M^*$ hold for all $\nu \geq \nu_1$ with some suitable $\nu_1 \geq 0$. Hence, the local corrections v_l^ν corresponding to $\mu_l^* \notin \mathcal{S}_j^\circ$ satisfy $v_l^\nu = \bar{v}_l^\nu = 0$, $\forall \nu \geq \nu_1$. In the remaining case $\mu_l^* \in \mathcal{S}_j^\circ$, the exact local corrections $\bar{v}_l^\nu = \bar{z}_l^\nu \mu_l^\nu$ tend to zero. Hence, we can find a $\nu_0 \geq \nu_1$ such that $\bar{z}_l^\nu \in I$, $\forall \nu \geq \nu_0$. Then it follows from (3.57) that $z_l^\nu = \bar{z}_l^\nu$, $\forall \nu \geq \nu_0$. This completes the proof. □

Assuming that (1.3.57) is non–degenerate, Theorem 2.9 states that the choice of different quasioptimal approximations does not change the asymptotic convergence properties of the resulting different iterative schemes. In particular, we obtain the same asymptotic convergence rates as for the extended relaxation method itself. For reasonable initial iterates, the asymptotic behavior may dominate the whole iteration process (cf. Kornhuber [82, 83]).

3 Monotone Multigrid Methods

In the preceding chapter, we introduced extended relaxation methods for the iterative solution of the discrete minimization problem (1.3.57). An extended relaxation method is characterized by the selection of a certain sequence of search directions for the successive minimization of the energy functional $\mathcal{J} + \phi_j$. Theorem 2.9 states that the global convergence together with the asymptotic convergence rates are preserved by quasioptimal monotone approximations of the local subproblems.

Extended relaxation methods were motivated by well–known multigrid methods for elliptic selfadjoint problems. In particular, the classical multigrid method with a Gauß–Seidel smoother can be regarded as an extended relaxation method induced by the multilevel nodal basis $\Lambda_{\mathcal{S}}$.

Monotone multigrid methods, based on the successive local minimization of energy, are intended to be a permanent extension of this classical approach to non–smooth minimization problems of the form (1.3.57). A monotone multigrid method should be globally convergent with asymptotic multigrid convergence rates and should allow an implementation as a usual V–cycle, requiring $\mathcal{O}(n_j)$ operations for each iteration step.

From the above considerations, the extended relaxation method induced by the multilevel nodal basis $\Lambda_{\mathcal{S}}$ is a natural candidate for a monotone multigrid method. However, for the nonlinear problem (1.3.57), this method can no longer be implemented with optimal numerical complexity.

As a consequence, we will derive a suitable linearization of the local subproblems. The basic idea of monotone multigrid methods is first to find out a neighborhood of the actual iterate in which the actual linearization is valid and then to constrain the coarse grid correction to this neighborhood. In this way, we end up with a sequence of local obstacle problems for the approximate corrections.

The resulting *standard monotone multigrid method* can be regarded as an extended underrelaxation induced by $\Lambda_{\mathcal{S}}$ and certain quasioptimal monotone approximations. Hence, it is globally convergent (cf. Theorem 2.9).

However, the underlying search directions taken from the multilevel nodal basis may be not well–suited to the phases of u_j. The resulting poor coarse grid transport usually causes unsatisfactory asymptotic convergence rates. By adapting the elements of $\Lambda_{\mathcal{S}}$ to the actual guess of the discrete phases in each iteration step, we obtain so–called *truncated monotone multigrid methods* with improved convergence properties. Note that similar ideas can be applied to selfadjoint elliptic problems with rapidly varying coefficients or complicated geometries. First steps in this direction were made by Kornhuber and Yserentant [88]. Truncations of the hierarchical basis were introduced by Hoppe and Kornhuber [75] in connection with obstacle problems. Related techniques were derived by Bank and Xu [12] and Hackbusch and Sauter [67] for the appropriate coarsening of a given mesh.

In the light of Theorem 2.9, both variants of the monotone multigrid method reduce asymptotically to a linear iteration on a reduced space. Hence, we can derive asymptotic estimates of the convergence rates by a suitable extension of the well–known linear theory.

Both the standard and the truncated monotone multigrid method can be implemented as a slight modification of the usual V–cycle. Some algorithmic remarks are made in the Sections 3.1.4 and 3.2.3 (see also Kornhuber [85]).

3.1 Standard Monotone Multigrid Methods

3.1.1 The Multilevel Nodal Basis

Let $\mathcal{T}_0$ be a triangulation of the polygonal domain Ω. The triangulation $\mathcal{T}_0$ is refined several times providing a sequence of triangulations $\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_j$. A triangle $t \in \mathcal{T}_k$ is refined either by subdividing it into four congruent subtriangles or by connecting one of its vertices with the midpoint of the opposite side. The first case is called *regular (red) refinement* and the resulting triangles are regular as well as the triangles of the initial triangulation $\mathcal{T}_0$. The second case is called *irregular (green) refinement* and results in two irregular triangles.

We say that the triangulations $\mathcal{T}_0, \ldots, \mathcal{T}_j$ are *nested*, if the global refinement process satisfies the following conditions (T1)–(T3).

Because new points should be generated only by regular refinement, we introduce the rule

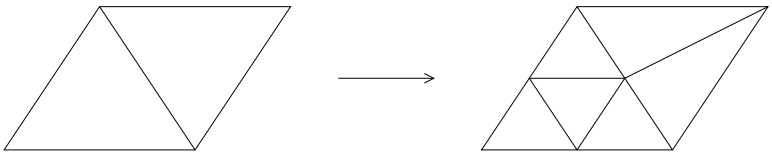

Figure 3.1 Regular refinement and irregular closure

(T1) Each vertex of $\mathcal{T}_{k+1}$ that does not belong to $\mathcal{T}_k$ is a vertex of a regular triangle.

Note that irregular refinement is potentially dangerous, because the interior angles are reduced. Hence we add the rule

(T2) Irregular triangles must not be further refined.

We say that a refined triangle is the *father* of the resulting triangles, which in turn are called *sons*. We define the *depth* of a given triangle $t \in \bigcup_{k=0}^{j} \mathcal{T}_k$ as the number of ancestors of t. Of course, the depth of all triangles $t \in \mathcal{T}_k$ is bounded by k. We have the final rule

(T3) Only triangles $t \in \mathcal{T}_k$ of depth k may be refined for the construction of $\mathcal{T}_{k+1}$, $0 \le k \le j$.

As a consequence of (T3), the whole sequence $\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_j$ can be uniquely reconstructed from the initial triangulation $\mathcal{T}_0$ and the final triangulation $\mathcal{T}_j$ alone. Note that the shape regularity of $\mathcal{T}_j$ is inherited from the shape regularity of $\mathcal{T}_0$. The conditions (T1)–(T3) are meanwhile standard in the field of multilevel methods (see e.g. Bank, Dupont and Yserentant [9], Bornemann, Erdmann and Kornhuber [27], Deuflhard, Leinen and Yserentant [45]).

The construction of the triangulations should be based on some adaptive strategy. Note that the sequence $\mathcal{T}_0, \ldots, \mathcal{T}_j$ does not necessarily reflect the underlying dynamic refinement process. We will come back to this point in Chapter 4.

A sequence $\mathcal{T}_0, \ldots, \mathcal{T}_j$ of nested triangulations gives rise to a sequence of *nested* finite element spaces

$$\mathcal{S}_0 \subset \mathcal{S}_1 \subset \ldots \subset \mathcal{S}_j.$$

Let $\Lambda_k = \{\lambda_p^{(k)} |\ p \in \mathcal{N}_k\}$ denote the set of nodal basis functions in $\mathcal{S}_k$, $k = 0, \ldots, j$. Collecting the $m_0 = n_0$ elements of Λ_0 and the m_k new basis functions on each level, we define the *multilevel nodal basis* $\Lambda_{\mathcal{S}}$ with $m_{\mathcal{S}} = m_0 + \ldots + m_j$ elements by

$$\Lambda_{\mathcal{S}} = \Lambda_0 \cup \bigcup_{k=1}^{j} \Lambda_k \setminus \Lambda_{k-1}.$$

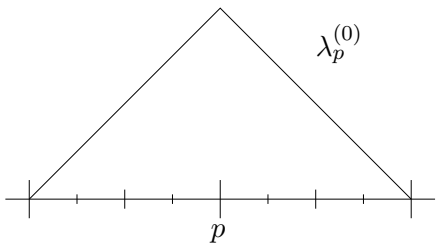

Figure 3.2 Low frequency multilevel nodal basis function $\lambda_p^{(0)} \in \Lambda_{\mathcal{S}}$

We use the canonical ordering of $\Lambda_{\mathcal{S}}$, which follows the refinement levels,

$$\Lambda_{\mathcal{S}} = (\lambda_{p_1}^{(j)}, \lambda_{p_2}^{(j)}, \ldots, \lambda_{p_{m_j}}^{(j)}, \ldots, \lambda_{p_1}^{(0)}, \ldots, \lambda_{p_{m_0}}^{(0)}). \tag{1.1}$$

In the case of elliptic selfadjoint problems, the classical multigrid V–cycle with Gauß–Seidel smoother and canonical restrictions and prolongations can be regarded as a linear extended relaxation method induced by the constant sequence $\Lambda_{\mathcal{S}}$ of search directions (cf. e.g. Xu [122]). Hence, the nonlinear extended relaxation method induced by $\Lambda_{\mathcal{S}}$ is a natural extension to nonlinear problems of the form (1.3.57).

However, in the case of non–uniform refinement, as considered here, the canonical ordering (1.1) of $\Lambda_{\mathcal{S}}$ contradicts the condition (M1) that each sequence of search directions should start with the elements $\lambda_p^{(j)}$ of the fine grid

nodal basis Λ_j (see Section 2.1.2, p. 51). Hence, we consider the extended relaxation method induced by

$$M^\nu = \Lambda = (\Lambda_j, \Lambda_{\mathcal{S}}) = (\lambda_{p_1}^{(j)}, \ldots, \lambda_{p_{n_j}}^{(j)}, \lambda_{n_j+1} \ldots, \lambda_m), \quad \forall \nu \geq 0. \quad (1.2)$$

The enumeration of the constant extension

$$M_c^\nu = \Lambda_{\mathcal{S}} = (\lambda_{n_j+1} \ldots, \lambda_m), \quad \forall \nu \geq 0,$$

with $m = n_j + m_{\mathcal{S}}$, follows the canonical ordering (1.1). It is clear that Λ also satisfies the conditions (M2) and (M3) stated in Section 2.3, p. 60.

Recall that the leading fine grid corrections in direction of $\lambda_{p_1}, \ldots, \lambda_{p_{n_j}}$ can be evaluated exactly by (2.1.20). Let $\lambda_l \in \Lambda_{\mathcal{S}} \cap \mathcal{S}_k$ with $k < j$. Then the corresponding coarse grid correction $\bar{v}_l$ cannot be computed without evaluating the intermediate iterate w_{l-1} at all nodes $p \in \text{int supp } \lambda_l$, because the subdifferential $\partial\phi_j(w_{l-1}+v)$ is nonlinear with respect to w_{l-1}. This leads to (at least) one additional prolongation for each local coarse grid correction. As a consequence, the number of operations for a global iteration step is no longer bounded by $\mathcal{O}(n_j)$. To preserve the optimal numerical complexity of the classical V–cycle, we will derive suitable approximations of the coarse grid problems. For simplicity, the fine grid corrections corresponding to $\lambda_l \in \Lambda_j \cap \Lambda_{\mathcal{S}}$ will not be treated separately.

3.1.2 Quasioptimal Approximations

Linear functions on subspaces $\mathcal{S}_k \subset \mathcal{S}_j$ can be represented by their values on the coarse grid basis functions $\lambda_p^{(k)} \in \Lambda_k$. This property gives rise to the canonical restrictions of the residual and of the stiffness matrix occurring in the implementation of linear subspace correction methods as a classical V–cycle.

Assume that the smoothed iterate $\bar{u}_j^\nu$ has been computed from a given iterate u_j^ν by the leading Gauß–Seidel relaxation. To take advantage of the simple representation of linear functions on the coarse grid spaces, we want to *restrict the subsequent intermediate iterates* w_l, $l = n_j + 1, \ldots, m$, *to a neighborhood of* $\bar{u}_j$, *on which the subdifferential* $\partial\phi_j(w)$ *is an affine function of* w. In doing so, we will exploit that ϕ_j is generated by a piecewise quadratic function Φ (see condition (V3)' stated in Section 1.2.1).

We define the closed, convex subset $\mathcal{K}_{\bar{u}_j^\nu} \subset \mathcal{S}_j$,

$$\mathcal{K}_{\bar{u}_j^\nu} = \{v \in \mathcal{S}_j | \, \underline{\varphi}_j^\nu(p) \le v(p) \le \overline{\varphi}_j^\nu(p), \ \forall p \in \mathcal{N}_j\},$$

where the obstacles $\underline{\varphi}_j^\nu, \overline{\varphi}_j^\nu \in \mathcal{S}_j$ are given by

$$\begin{aligned} \underline{\varphi}_j^\nu(p) = \theta_i, \ \ \overline{\varphi}_j^\nu(p) = \theta_{i+1}, \ \text{ if } \ \bar{u}_j^\nu(p) \in (\theta_i, \theta_{i+1}) \\ \underline{\varphi}_j^\nu(p) = \overline{\varphi}_j^\nu(p) = \theta_i, \qquad \text{if} \qquad \bar{u}_j^\nu(p) = \theta_i, \end{aligned} \tag{1.3}$$

for $i = 0, \ldots, N+1$ and all $p \in \mathcal{N}_j$. As usual, the index ν will frequently be skipped in the sequel.

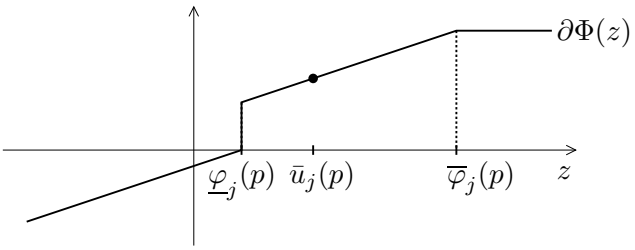

Figure 3.3 Local linearization

By construction of the obstacles $\underline{\varphi}_j$ and $\overline{\varphi}_j$, the functional ϕ_j on $\mathcal{K}_{\bar{u}_j}$ can be rewritten in the form

$$\phi_j(w) = \tfrac{1}{2} b_{\bar{u}_j}(w, w) - f_{\bar{u}_j}(w) + \text{const.}, \quad \forall w \in \mathcal{K}_{\bar{u}_j}, \tag{1.4}$$

giving

$$\partial\phi_j(w)(v) = b_{\bar{u}_j}(w, v) - f_{\bar{u}_j}(v), \quad \forall w \in \mathcal{K}_{\bar{u}_j}, \ \ \forall v \in \mathcal{S}_j. \tag{1.5}$$

The bilinear form $b_{\bar{u}_j}(\cdot, \cdot)$ and the functional $f_{\bar{u}_j}$ on $\mathcal{S}_j$ are defined by (2.3.50) and (2.3.51), respectively, replacing u_j by $\bar{u}_j$. Observe that the underlying approximation of the discrete phases

$$\mathcal{N}_j = \mathcal{N}_j^\bullet(\bar{u}_j) \cup \bigcup_{i=0}^{N} \mathcal{N}_j^i(\bar{u}_j) \tag{1.6}$$

is fixed by the fine grid relaxation. For the definition of discrete phases, we refer to (2.3.40) and (2.3.41).

It is trivial that $\bar{u}_j \in \mathcal{K}_{\bar{u}_j}$. As the local linearization (1.5) is only valid in the neighborhood $\mathcal{K}_{\bar{u}_j}$ of $\bar{u}_j$, we impose the condition

$$v_l^* \in \mathcal{D}_l^* \subset V_l, \quad \forall l = n_j + 1, \ldots, m, \tag{1.7}$$

on the subsequent corrections v_l^*. The subsets $\mathcal{D}_l^*$,

$$\mathcal{D}_l^* = \{v \in V_l \mid \underline{\varphi}_j(p) - w_{l-1}(p) \leq v(p) \leq \overline{\varphi}_j(p) - w_{l-1}(p),\ \forall p \in \mathcal{N}_j\},$$

are chosen in such a way that $w_l \in \mathcal{K}_{\bar{u}_j}$, $l = n_j + 1, \ldots, m$. Equivalently, *the coarse grid corrections must not cause a change of phase*. In particular, the values at the critical points $p \in \mathcal{N}_j^\bullet(\bar{u}_j)$ remain invariant.

To satisfy the condition (1.7), we approximate the local subproblems (2.1.22) by the constrained minimization problems

$$\begin{aligned} v_l^* \in \mathcal{D}_l^* : \ & \mathcal{J}(w_{l-1} + v_l^*) + \phi_j(w_{l-1} + v_l^*) \\ & \leq \mathcal{J}(w_{l-1} + v) + \phi_j(w_{l-1} + v), \ \ \forall v \in \mathcal{D}_l^*. \end{aligned} \tag{1.8}$$

By virtue of the representation (1.5), the local obstacle problems (1.8) can be reformulated as the variational inequalities

$$\begin{aligned} v_l^* \in \mathcal{D}_l^* : \ & a_{\bar{u}_j}(v_l^*, v - v_l^*) \\ & \geq \ell_{\bar{u}_j}(v - v_l^*) - a_{\bar{u}_j}(w_{l-1}, v - v_l^*), \ \ \forall v \in \mathcal{D}_l^*, \end{aligned} \tag{1.9}$$

where the bilinear form $a_{\bar{u}_j}(\cdot,\cdot)$ and the functional $\ell_{\bar{u}_j}$ are defined in analogy to (2.3.52).

It will follow from a later result that the approximate subproblems (1.9) correspond to certain quasioptimal approximations. However, the solution of (1.9) still cannot be obtained in an efficient way, because the definition of the constraints $\mathcal{D}_l^*$ makes use of the values $w_{l-1}(p)$, $p \in \mathcal{N}_j$. Recall that the evaluation of $w_l(p)$ for *all* $p \in \mathcal{N}_j$ *and all* $l = n_j + 1, \ldots, m$ spoils the optimal numerical complexity.

Hence, the subproblems (1.9) are replaced by further approximations

$$\begin{aligned} v_l \in \mathcal{D}_l : \ & a_{\bar{u}_j}(v_l, v - v_l) \\ & \geq \ell_{\bar{u}_j}(v - v_l) - a_{\bar{u}_j}(w_{l-1}, v - v_l), \ \ \forall v \in \mathcal{D}_l, \end{aligned} \tag{1.10}$$

where the closed, convex subsets $\mathcal{D}_l \subset V_l$,

$$\mathcal{D}_l = \{v \in V_l \mid \underline{\psi}_l(p) \le v(p) \le \overline{\psi}_l(p),\ \forall p \in \mathcal{N}_j\},$$

are intended to approximate the constraints $\mathcal{D}_l^*$. Consequently, the *local obstacles* $\underline{\psi}_l, \overline{\psi}_l \in V_l$ should approximate the defect obstacles $\underline{\varphi}_j - w_{l-1}$ and $\overline{\varphi} - w_{l-1}$, respectively. Observe that the condition $\underline{\psi}_l, \overline{\psi}_l \in V_l$ allows us to check the constraints without visiting the fine grid.

To give a reinterpretation of the approximate subproblems (1.10) in terms of quasioptimal approximations (cf. Definition 2.8), we will reformulate (1.10) as variational inclusions of the form (2.2.38). For this reason, we define the scalar, convex functions Ψ_l for all $l = n_j + 1, \ldots, m$ by

$$\Psi_l(z) = \phi_j(w_{l-1} + z\lambda_l) + \chi_l(z), \quad \forall z \in \mathbb{R}, \tag{1.11}$$

where χ_l denotes the characteristic function of $I_l = \{z \in \mathbb{R} \mid z\lambda_l \in \mathcal{D}_l\} \subset \mathbb{R}$. Then it is easily checked that (1.10) can be rewritten as

$$z_l \in I_l: \quad 0 \in a_{ll} z_l - r_l + \partial\Psi_l(z_l) \tag{1.12}$$

and $v_l = z_l\lambda_l$. Recall the notation $a_{ll} = a(\lambda_l, \lambda_l)$ and $r_l = \ell(\lambda_l) - a(w_{l-1}, \lambda_l)$. The local obstacles $(\underline{\psi}_l^\nu)_{\nu\ge0}$, $(\overline{\psi}_l^\nu)_{\nu\ge0}$, are called *quasioptimal*, if $(\partial\Psi_l^\nu)_{\nu\ge0}$ is a quasioptimal approximation of $(\partial\Phi_l^\nu)_{\nu\ge0}$. The following lemma gives sufficient criteria for the quasioptimality.

Lemma 3.1 *Assume that for all $l = n_j + 1, \ldots, m$ and all $\nu \ge 0$ the local obstacles $\underline{\psi}_l^\nu$ and $\overline{\psi}_l^\nu$ are continuous functions of $w_{n_j}^\nu, \ldots, w_{l-1}^\nu$,*

$$\underline{\psi}_l^\nu = \underline{\psi}_l(w_{n_j}^\nu, \ldots, w_{l-1}^\nu),\ \ \overline{\psi}_l^\nu = \overline{\psi}_l(w_{n_j}^\nu, \ldots, w_{l-1}^\nu),$$

which are monotone in the sense that $w_{n_j}^\nu, \ldots, w_{l-1}^\nu \in \mathcal{K}_{\bar{u}_j^\nu}$ implies

$$\underline{\varphi}_j^\nu(p) - w_{l-1}^\nu(p) \le \underline{\psi}_l^\nu(p) \le 0 \le \overline{\psi}_l^\nu(p) \le \overline{\varphi}_j^\nu(p) - w_{l-1}^\nu(p) \tag{1.13}$$

for all $p \in$ int supp λ_l.

Assume further that for $l = n_j + 1, \ldots, m$ the functions $\underline{\psi}_l^* = \underline{\psi}_l(u_j, \ldots, u_j)$ and $\overline{\psi}_l^* = \overline{\psi}_l(u_j, \ldots, u_j)$ satisfy

$$\underline{\psi}_l^*(p) < 0 < \overline{\psi}_l^*(p), \quad \forall p \in \text{int supp } \lambda_l, \tag{1.14}$$

if $\lambda_l(p) = 0$ *holds for all* $p \in \mathcal{N}_j^\bullet(u_j)$.

Then the local obstacles $(\underline{\psi}_l^\nu)_{\nu\geq 0}$, $(\overline{\psi}_l^\nu)_{\nu\geq 0}$ *are quasioptimal.*

Proof. Consider some fixed $l = n_j+1, \ldots, m$. It is clear from $\bar{u}_j = w_{n_j} \in \mathcal{K}_{\bar{u}_j}$ and (1.13) that $0 \in I_l$. Now the monotonicity (2.1.25) follows from

$$\Psi_l(z) = \Phi_l(z) + \chi_l(z), \quad \forall z \in \mathbb{R}, \tag{1.15}$$

and simple arguments from convex analysis.

Assume that the intermediate iterates $(w_l^\nu)_{\nu\geq 0}$ and their phases are convergent in the sense of (2.3.43). Choose $\nu_0 \geq 0$ such that $\mathcal{N}_j^\bullet(w_l^\nu) = \mathcal{N}_j^\bullet(u_j)$ holds for all $l = 1, \ldots, m$ and all $\nu \geq \nu_0$. As $\underline{\psi}_l^\nu$, $\overline{\psi}_l^\nu$ depend continuously on $w_{n_j}, \ldots, w_m$ the convergence of $(w_l^\nu)_{\nu\geq 0}$ implies

$$\underline{\psi}_l^\nu(p) \to \underline{\psi}_l^*(p), \;\; \overline{\psi}_l^\nu(p) \to \overline{\psi}_l^*(p), \qquad \nu \to \infty. \tag{1.16}$$

The convergence is uniform with respect to $l = n_j + 1, \ldots, m$ and $p \in \mathcal{N}_j$. It is easily seen that the inequalities in (1.14) also hold uniformly with respect to l and p. Hence, we can find a positive number $\varepsilon > 0$ and an index $\nu_1 \geq \nu_0$ such that

$$\underline{\psi}_l^\nu(p) \leq -\varepsilon < 0 < \varepsilon \leq \overline{\psi}_l^\nu(p), \quad \forall p \in \mathcal{N}_j \cap \text{int supp } \lambda_l, \tag{1.17}$$

is valid for all $\nu \geq \nu_1$, if λ_l is vanishing on $\mathcal{N}_j^\bullet(u_j)$. Setting $I = (-\varepsilon, \varepsilon)$, we clearly have $0 \in I \subset I_l$ and

$$\partial\Psi_l^\nu(z) = \partial\Phi_l^\nu(z), \quad \forall z \in I, \quad \forall \nu \geq \nu_1, \tag{1.18}$$

if $\lambda_l(p) = 0$ holds for all $p \in \mathcal{N}_j^\bullet(u_j)$. This completes the proof. □

We emphasize that no particular properties of the multilevel nodal basis entered our preceding considerations. In fact, the construction of quasioptimal approximations by suitable obstacle problems of the form (1.10) can be generalized to any regular sequence $(M^\nu)_{\nu\geq 0}$ of search directions.

3.1.3 Quasioptimal Restrictions

To complete the construction of a monotone multigrid method, we now derive quasioptimal local obstacles $\underline{\psi}_l$ and $\overline{\psi}_l$ for $l = n_j + 1, \ldots, m$. For symmetry reasons, it is sufficient to consider only the upper obstacles $\overline{\psi}_l$. The construction relies on suitable successive restrictions of the upper defect obstacles $\overline{\varphi}_j - w_l^\nu$.

We say that $\lambda_l \in \Lambda$ is *on level* k, if $\lambda_l \in \Lambda_k$. To identify the supporting points and the levels of $\lambda_l \in \Lambda_{\mathcal{S}}$, we will use the notation

$$\lambda_{l_{ik}} = \lambda_{p_i}^{(k)}, \qquad i = 1, \ldots, m_k, \quad k = 0, \ldots, j.$$

Then the correction

$$v^{(k)} = v_{p_1}^{(k)} + \ldots + v_{p_{m_k}}^{(k)} \tag{1.19}$$

is the sum of all local corrections $v_{l_{ik}} = v_{p_i}^{(k)}$ in direction of the basis functions $\lambda_{l_{ik}} = \lambda_{p_i}^{(k)}$ on level k. Recall that the local corrections v_l are obtained from the local obstacle problems (1.10).

Lemma 3.2 *Assume that the mappings* $R_{k+1}^k : \mathcal{S}_{k+1} \to \mathcal{S}_k$, $k = j-1, \ldots, 0$, *are continuous and that the conditions*

$$R_{k+1}^k v(p) \le v(p), \quad \forall p \in \mathcal{N}_{k+1}, \tag{1.20}$$

and

$$\min\{v(q) \mid q \in \mathcal{N}_{k+1} \cap \text{ int supp } \lambda_p^{(k)}\} \le R_{k+1}^k v(p), \ \ \forall p \in \mathcal{N}_k, \tag{1.21}$$

hold for all non–negative $v \in \mathcal{S}_{k+1}$. *Then, for a given iterate* u_j^ν *and the initial defect obstacle* $\overline{\psi}^{(j)} = \overline{\varphi}_j - u_j^\nu$, *the recursive restriction*

$$\overline{\psi}^{(k)} = R_{k+1}^k(\overline{\psi}^{(k+1)} - v^{(k+1)}), \ \ k = j-1, \ldots, 0, \tag{1.22}$$

inductively provides quasioptimal local obstacles $\overline{\psi}_l \in V_l$ *by the definition*

$$\overline{\psi}_{l_{ik}} = \overline{\psi}^{(k)}(p_i)\lambda_{p_i}^{(k)}, \quad i = 1, \ldots, m_k, \ \ k = j-1, \ldots, 0. \tag{1.23}$$

Proof. Observe that we have

$$\psi^{(j)} - v^{(j)} = \overline{\varphi}_j - (u_j^\nu + v^{(j)}) \geq 0,$$

because $\overline{\varphi}_j - \bar{u}_j^\nu \geq 0$ holds by construction of $\overline{\varphi}_j$, and this property is preserved by the subsequent local fine grid corrections computed from the local obstacle problems (1.8).

Now we can inductively apply the condition (1.20) to show the monotonicity (1.13) of the local upper obstacles ψ_l defined in (1.23). The property (1.14) follows in a similar way from the condition (1.21), and the definition of $\overline{\varphi}_j$ together with the continuity of the restrictions R_{k+1}^k implies that the local obstacles are continuous functions of the preceding intermediate iterates. □

We are left with the problem of constructing *quasioptimal restriction operators* $R_{k+1}^k : \mathcal{S}_{k+1} \to \mathcal{S}_k$ satisfying the assumptions of Lemma 3.2.

It is easily seen that the restrictions $r_{k+1}^k : \mathcal{S}_{k+1} \to \mathcal{S}_k$, $k = 0, \ldots, j-1$,

$$r_{k+1}^k v(p) = \min\{v(q) \mid q \in \mathcal{N}_{k+1} \cap \text{ int supp } \lambda_p^{(k)}\}, \quad \forall p \in \mathcal{N}_k, \quad (1.24)$$

proposed by Mandel [93, 92], are quasioptimal in this sense. Though the definition (1.24) looks quite natural in the light of condition (1.21), it does not take advantage of the fact that the arguments $v \in \mathcal{S}_{k+1}$ of r_{k+1}^k are piecewise linear on $\mathcal{T}_k$. As a consequence, the resulting local constraints are too pessimistic as compared to the quasioptimal restrictions R_{k+1}^k which we will derive now.

For some fixed k, $0 \leq k \leq j-1$, let $\mathcal{E}_k' \subset \mathcal{E}_k$ denote the subset of bisected edges, and $p_e \in \mathcal{N}_{k+1}$ is the midpoint of $e \in \mathcal{E}_k'$. Selecting a certain order $\mathcal{E}_k' = (e_1, \ldots, e_s)$, we define the restriction operator $R_{k+1}^k : \mathcal{S}_{k+1} \to \mathcal{S}_k$ according to

$$R_{k+1}^k v = I_{\mathcal{S}_k} \circ R_{e_s} \circ \ldots \circ R_{e_1} v. \quad (1.25)$$

Here $I_{\mathcal{S}_k}$ denotes the $\mathcal{S}_k$–interpolation, and for each $e \in \mathcal{E}_k'$ the operator $R_e : \mathcal{S}_{k+1} \to \mathcal{S}_{k+1}$ is of the form

$$R_e v = v + v_1 \lambda_{p_1}^{(k+1)} + v_2 \lambda_{p_2}^{(k+1)}, \quad (1.26)$$

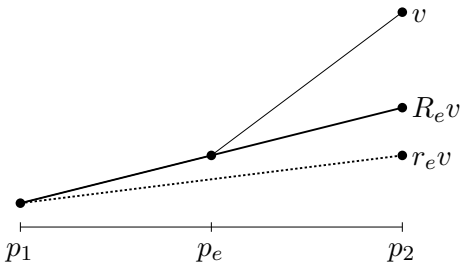

Figure 3.4 Local monotone restriction

with $p_1, p_2 \in \mathcal{N}_k$ denoting the vertices of $e = (p_1, p_2) \in \mathcal{E}'_k$. The scalars $v_1, v_2 \in \mathbb{R}$ in (1.26) are chosen such that

$$R_e v(p) \leq v(p), \quad p = p_1, p_e, p_2.$$

In particular, we set $v_1 = 0$, if $v(p_1) \leq v(p_e)$ or if $\frac{1}{2}(v(p_1) + v(p_2)) \leq v(p_e)$. In the remaining case, v_1 is determined by

$$v_1 = \begin{cases} 2v(p_e) - v(p_1) - v(p_2), & \text{if } v(p_2) \leq v(p_e) \leq v(p_1) \\ v(p_e) - v(p_1), & \text{if } v(p_e) \leq \min\{v(p_1), v(p_2)\} \end{cases}.$$

The value of v_2 is obtained in a symmetrical way.

It can be checked by elementary considerations that, for any enumeration of $\mathcal{E}'_k$, the definition (1.25) provides a quasioptimal restriction operator R^k_{k+1}. Moreover, decomposing r^k_{k+1} in local restriction operators r_e, $e \in \mathcal{E}'_k$, in analogy to (1.25), it can be shown that

$$r^k_{k+1} v \leq R^k_{k+1} v \tag{1.27}$$

holds for all non–negative $v \in \mathcal{S}_{k+1}$. This is illustrated in Figure 3.4. Hence, using R^k_{k+1} instead of r^k_{k+1}, we can expect less damping of the coarse–grid corrections, providing faster convergence of the corresponding algorithm. On the other hand, if the underlying problem is non–degenerate, then we know from the preceding chapter that the asymptotic behavior of both methods must be the same. This partly heuristic reasoning is strengthened by numerical experiments (cf. Kornhuber [82, 83]).

3.1.4 A Standard Multigrid V–Cycle

The extended underrelaxation provided by the local subproblems (1.10) with upper local obstacles $(\overline{\psi}_l^\nu)_{\nu\geq 0}$ generated according to Lemma 3.2 by the quasioptimal upper restrictions $\overline{R}_{k+1}^k = R_{k+1}^k$ and with lower local obstacles $(\underline{\psi}_l^\nu)_{\nu\geq 0}$ generated in a similar way by the lower counterparts $\underline{R}_{k+1}^k$ is called the *standard monotone multigrid method induced by* Λ. Other standard monotone multigrid methods are characterized by other *constant* sequences of search directions which contain the multilevel nodal basis $\Lambda_{\mathcal{S}}$ and start with the nodal basis Λ_j (cf. condition (M1), p. 51).

The standard monotone multigrid method induced by Λ can be implemented as a classical V–cycle.

Algorithm 3.1 (Standard Monotone Multigrid Method)

given iterate: u_j^ν

fine grid smoothing: $\bar{u}_j^\nu := \mathcal{M}_j(u_j^\nu)$

local linearization: $a_{\bar{u}_j^\nu} := a + b_{\bar{u}_j^\nu}, \quad \ell_{\bar{u}_j^\nu} := \ell + f_{\bar{u}_j^\nu}$

coarse grid correction:

initialize:

bilinear form and residual: $a^{(j)} := a_{\bar{u}_j^\nu}, \quad r^{(j)} := \ell_{\bar{u}_j^\nu} - a_{\bar{u}_j^\nu}(\bar{u}_j^\nu, \cdot)$

defect obstacles: $\underline{\psi}^{(j)} := \underline{\varphi}_j^\nu - \bar{u}_j^\nu, \quad \overline{\psi}^{(j)} := \overline{\varphi}_j^\nu - \bar{u}_j^\nu$

global correction: $v_j^\nu := 0$

for $k = j-1$ *step* -1 *until* 0 *do*

canonical restrictions: $a^{(k)} := a^{(k+1)}|_{\mathcal{S}_k \times \mathcal{S}_k}, \quad r^{(k)} := r^{(k+1)}|_{\mathcal{S}_k}$

quasioptimal restrictions: $\underline{\psi}^{(k)} := \underline{R}_{k+1}^k \underline{\psi}^{(k+1)}, \quad \overline{\psi}^{(k)} := \overline{R}_{k+1}^k \overline{\psi}^{(k+1)}$

coarse grid smoothing: $v^{(k)} := \bar{\mathcal{M}}_k(a^{(k)}, r^{(k)}, \underline{\psi}^{(k)}, \overline{\psi}^{(k)})(0)$

update:

residual: $r^{(k)} := r^{(k)} - a^{(k)}(v^{(k)}, \cdot)$

defect obstacles: $\underline{\psi}^{(k)} := \underline{\psi}^{(k)} - v^{(k)}, \quad \overline{\psi}^{(k)} := \overline{\psi}^{(k)} - v^{(k)}$

for $k = 0$ *step* 1 *until* $j-1$ *do*

canonical interpolation: $v_j^\nu := v_j^\nu + v^{(k)}$

new iterate: $u_j^{\nu+1} := \bar{u}_j^\nu + v_j^\nu$

Recall that $\mathcal{M}_j$ stands for one step of the nonlinear Gauß–Seidel relaxation (2.1.3). The resulting smoothed iterate $\bar{u}_j^\nu$ determines the actual fine grid obstacles $\underline{\varphi}_j^\nu$ and $\overline{\varphi}_j^\nu$ according to (1.3). These constraints define the neighborhood $\mathcal{K}_{\bar{u}_j^\nu}$ of $\bar{u}_j^\nu$ in which the local linearization (1.5) is valid. The bilinear form $a^{(k)}(\cdot,\cdot)$ and the actual residual $r^{(k)}$ can be applied directly to the elements of the subspaces $\mathcal{S}_k$ which gives the canonical restriction. In (1.25), we have just defined the (quasioptimal) restriction $\overline{R}_{k+1}^k$ of the upper defect obstacle $\overline{\psi}^{(k)}$. The restriction $\underline{R}_{k+1}^k$ of the lower counterpart $\underline{\psi}^{(k)}$ is performed in a symmetrical way.

The evaluation of the correction $v^{(k)}$ (see 1.19) from the approximate local coarse grid problems (1.10) can be rewritten as a projected Gauß–Seidel iteration on level k. For given bilinear form a, right–hand side r, and obstacles $\underline{\psi}$, $\overline{\psi}$, the corresponding iteration operator is denoted by $\bar{\mathcal{M}}_k(a, r, \underline{\psi}, \overline{\psi})$.

The implementation of Algorithm 3.1 requires its reformulation in terms of vectors and matrices. To this end, each of the nodal basis functions $\lambda_p^{(k)}$ is identified with a unit vector of the corresponding Euclidean space. In the linear selfadjoint case, the resulting transformation of the algorithm is described for example by Hackbusch [65] or Braess [30]. A corresponding linear multigrid code can be extended to an implementation of Algorithm 3.1 simply by using nonlinear or projected Gauß–Seidel smoothers and adding the quasioptimal restrictions.

As a consequence of Theorem 2.9, we have the following

Proposition 3.3 *The standard monotone multigrid method induced by* Λ *is globally convergent. If the discrete problem (1.3.57) is non–degenerate in the sense of (2.3.42), then the standard monotone multigrid method reduces asymptotically to the linear extended relaxation method induced by* Λ°*,*

$$\Lambda^\circ = \Lambda \cap \mathcal{S}_j^\circ, \tag{1.28}$$

for the corresponding linear reduced problem (2.3.53). The ordering of the subset $\Lambda^\circ \subset \Lambda$ *is inherited from* Λ*.*

Observe that all elements $\lambda_l \in \Lambda$ with $\lambda_l \notin \Lambda^\circ$, i.e. with the property $\operatorname{int}\operatorname{supp} \lambda_l \cap \mathcal{N}_j^\bullet(u_j) \neq \emptyset$, asymptotically do not contribute to the coarse grid correction. This leads to deteriorated asymptotic convergence rates as compared to classical linear multigrid methods.

3.2 Truncated Monotone Multigrid Methods

In the special case of obstacle problems, the standard monotone multigrid method presented above reduces to a variant of the multigrid method of Mandel [93, 92]. It turned out in numerical experiments that this method usually converges not as fast as the algorithm of Brandt and Cryer [33]. However, for the latter method there is no convergence proof. It was shown in a recent paper by Kornhuber [82] that the combination of satisfying theoretical *and* numerical properties can be achieved by adapting the multilevel nodal basis to the discrete phases by suitable truncation.

3.2.1 Truncation of the Multilevel Nodal Basis

The overall convergence of a monotone multigrid method is frequently dominated by the asymptotic convergence rates (cf. Kornhuber [82, 83]). To improve the asymptotic convergence by improved coarse–grid transport, we will now extend the set Λ° by suitable *truncation* of the multilevel nodal basis functions $\lambda_l \in \Lambda_{\mathcal{S}}$.

Assume that a smoothed iterate $\bar{u}_j^\nu$ has been computed by the fine grid relaxation, providing the set of critical nodes $\mathcal{N}_j^\bullet(\bar{u}_j^\nu)$. We define

$$\tilde{\lambda}_p^{(k)} = T_{j,k}^\nu \lambda_p^{(k)}, \quad \forall p \in \mathcal{N}_k, \tag{2.29}$$

using the truncation operators $T_{j,k}^\nu$, $k = 0, \ldots, j$,

$$T_{j,k}^\nu = I_{\mathcal{S}_j^\nu} \circ \ldots \circ I_{\mathcal{S}_k^\nu}. \tag{2.30}$$

Here $I_{\mathcal{S}_k^\nu} : \mathcal{S}_j \to \mathcal{S}_k^\nu$ denotes the $\mathcal{S}_k^\nu$–interpolation and the spaces $\mathcal{S}_k^\nu \subset \mathcal{S}_k$,

$$\mathcal{S}_k^\nu = \{v \in \mathcal{S}_k \mid v(p) = 0,\ \forall p \in \mathcal{N}_k^\nu\} \subset \mathcal{S}_k, \tag{2.31}$$

are the reduced subspaces with respect to $\mathcal{N}_k^\nu = \mathcal{N}_k \cap \mathcal{N}_j^\bullet(\bar{u}_j^\nu)$, $k = 0, \ldots, j$. A truncated nodal basis functions may have a rather strange shape. In particular, their support does not need to be connected. For a one–dimensional analogue this is illustrated in Figure 3.5 where the two critical nodes are marked by dots.

Note that we have

$$\tilde{\lambda}_p^{(k)} = 0, \quad \forall p \in \mathcal{N}_k^\nu.$$

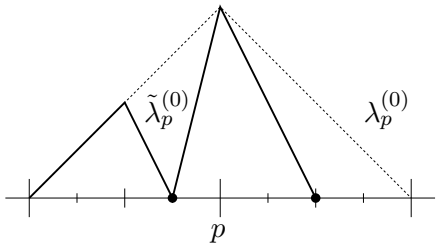

Figure 3.5 Truncated nodal basis function $\tilde{\lambda}_p^{(0)} \in \tilde{\Lambda}_{\mathcal{S}}^{\nu}$

The ordering of the remaining functions $\tilde{\lambda}_{l_{ik}}^{\nu} := \tilde{\lambda}_{p_i}^{(k)}$, $p_i \in \mathcal{N}_k \setminus \mathcal{N}_k^{\nu}$, is inherited from $\Lambda_{\mathcal{S}}$ providing the *variable* extension

$$M_c^{\nu} = \tilde{\Lambda}_{\mathcal{S}}^{\nu} = (\tilde{\lambda}_{n_j+1}^{\nu}, \ldots, \tilde{\lambda}_{\tilde{m}^{\nu}}^{\nu}), \quad \forall \nu \geq 0. \tag{2.32}$$

$\tilde{\Lambda}_{\mathcal{S}}^{\nu}$ is called *truncated multilevel nodal basis*. The extension (2.32) leads to

$$M^{\nu} = \tilde{\Lambda}^{\nu} = (\Lambda_j, \tilde{\Lambda}_{\mathcal{S}}^{\nu}), \quad \forall \nu \geq 0. \tag{2.33}$$

Because there is only a finite number of possible truncations on a fixed grid $\mathcal{T}_j$, the sequence $(\tilde{\Lambda}^{\nu})_{\nu \geq 0}$ satisfies condition (M2) stated in Section 2.3, p. 60. Obviously, $\tilde{\Lambda}^{\nu}$ depends only on the actual set of critical nodes $\mathcal{N}_j^{\bullet}(\bar{u}_j^{\nu})$ so that condition (M3) is also fulfilled. In particular, the reduced set $\tilde{\Lambda}^{\circ} \subset \mathcal{S}_j^{\circ}$ is determined by $\mathcal{N}_j^{\bullet}(u_j)$. It is easily checked that

$$\tilde{\lambda}_p^{(k)} = \lambda_p^{(k)}, \quad \forall \lambda_p^{(k)} \in \Lambda^{\circ},$$

so that $\Lambda^{\circ} \subset \tilde{\Lambda}^{\circ}$. Hence, we can hope for improved asymptotic convergence rates of the extended relaxation method induced by $(\tilde{\Lambda}^{\nu})_{\nu \geq 0}$ as compared to the standard case.

3.2.2 Quasioptimal Approximations and Restrictions

It follows by the same reasoning as above that the extended relaxation method induced by $(\tilde{\Lambda}^\nu)_{\nu\geq 0}$ cannot be implemented with optimal numerical complexity.

Hence, we replace the exact local subproblems (2.1.22) for the intermediate iterates w_l by obstacle problems of the form

$$\begin{aligned} v_l \in \mathcal{D}_l : \quad & a_{\bar{u}_j}(v_l, v - v_l) \\ & \geq \ell_{\bar{u}_j}(v - v_l) - a_{\bar{u}_j}(w_{l-1}, v - v_l), \quad \forall v \in \mathcal{D}_l, \end{aligned} \tag{2.34}$$

with closed, convex subsets $\mathcal{D}_l \subset V_l = \text{span}\,\{\tilde{\lambda}_l\}$,

$$\mathcal{D}_l = \{v \in V_l \mid \underline{\psi}_l(p) \leq v(p) \leq \overline{\psi}_l(p),\ \forall p \in \mathcal{N}_j\},$$

based on the local obstacles $\underline{\psi}_l$, $\overline{\psi}_l \in V_l$, $l = n_j + 1, \ldots, \tilde{m}$.

Again, the subproblems (2.34) correspond to quasioptimal approximations if the local obstacles $\underline{\psi}_l$ and $\overline{\psi}_l$ are properly chosen. In particular, Lemma 3.1 can be literally extended to the actual case of truncated search directions.

Utilizing a corresponding variant of Lemma 3.2, we can derive quasioptimal upper obstacles $\overline{\psi}_l$ by recursive restriction,

$$\overline{\psi}_l = \tilde{R}^k_{k+1}(\overline{\psi}^{(k+1)} - v^{(k+1)}), \quad k = j-1, \ldots, 0, \tag{2.35}$$

of the initial upper defect obstacle $\overline{\psi}(j) = \overline{\varphi}_j - u^\nu_j$ and the definition

$$\overline{\psi}_{l_{ik}} = \overline{\psi}^{(k)}(p_i)\tilde{\lambda}^{(k)}_{p_i}, \quad i = 1, \ldots, m_k, \ \ k = j-1, \ldots, 0. \tag{2.36}$$

Appropriate restriction operators $\tilde{R}^k_{k+1} : \mathcal{S}_{k+1} \to \mathcal{S}_k$, $k = j-1, \ldots, 0$, are obtained by a slight modification of the restrictions R^k_{k+1}. More precisely, we set

$$\tilde{R}^k_{k+1} v = I_{\mathcal{S}_k} \circ \tilde{R}_{e_s} \circ \ldots \circ \tilde{R}_{e_1} v. \tag{2.37}$$

To obtain the local restriction operator $\tilde{R}_e$ for each $e \in \mathcal{E}'_k$, we formally set $v(p_e) = \infty$ if $p_e \in \mathcal{N}^\nu_{k+1}$ and then compute the coefficients v_1 and v_2 appearing in (1.26) in the same way as before.

Quasioptimal lower obstacles $\underline{\psi}_l$, $l = n_j + 1, \ldots, \tilde{m}$, can be derived by symmetry arguments.

3.2.3 A Truncated Multigrid V–Cycle

The extended underrelaxation provided by the local subproblems (2.34) with upper local obstacles $(\overline{\psi}_l^\nu)_{\nu\geq 0}$ generated according to (2.36) by the quasioptimal restriction operators $\tilde{R}_{k+1}^k$ defined in (2.37) and with lower local obstacles $(\underline{\psi}_l^\nu)_{\nu\geq 0}$ generated in a symmetrical way is called the *truncated monotone multigrid method induced by* $(\tilde{\Lambda}^\nu)_{\nu\geq 0}$. Other truncated monotone multigrid methods result from the truncation of other constant sequences of search directions which contain the multilevel nodal basis $\Lambda_{\mathcal{S}}$ and start with the nodal basis Λ_j (cf. condition (M1), p. 51).

Truncated monotone multigrid methods can be implemented as a slight modification of the related standard version. In fact, we only have to modify the restriction of the stiffness matrix, of the residual and of the defect obstacles in such a way that there is *no contribution from values at the actual critical nodes* $p \in \mathcal{N}_j^\bullet(\bar{u}_j^\nu)$. Recall the definition of $\mathcal{N}_j^\bullet(\bar{u}_j^\nu)$ in (2.3.41).

Modifications of Algorithm 3.1 (Truncated Monotone Multigrid Method)

modified restrictions of the stiffness matrix and of the residual:

set all entries from the actual critical nodes $\mathcal{N}_j^\bullet(\bar{u}_j^\nu)$ *to zero*

modified quasioptimal restrictions of the upper (lower) defect obstacle:

set all entries from the actual critical nodes $\mathcal{N}_j^\bullet(\bar{u}_j^\nu)$ *to* ∞ $(-\infty)$

modified prolongations of the corrections:

set all prolongations to the actual critical nodes $\mathcal{N}_j^\bullet(\bar{u}_j^\nu)$ *to zero*

For the residual (but *not* for the stiffness matrix) such modified restrictions have been used for quite a while in connection with obstacle problems (cf. Brandt and Cryer [33], Hackbusch and Mittelmann [66]).

The following convergence result is a special case of Theorem 2.9.

Proposition 3.4 *The truncated monotone multigrid method induced by the sequence* $(\tilde{\Lambda}^\nu)_{\nu\geq 0}$ *is globally convergent. If the discrete problem (1.3.57) is non–degenerate in the sense of (2.3.42), then the truncated monotone multigrid method reduces asymptotically to the linear extended relaxation method induced by* $\tilde{\Lambda}^\circ$ *for the corresponding linear reduced problem (2.3.53).*

3.3 Asymptotic Convergence Rates

In this section, we will derive asymptotic estimates of the convergence rates of some standard and truncated monotone multigrid methods. By virtue of the asymptotic properties stated in Proposition 3.3 and Proposition 3.4, we only have to consider related successive subspace correction methods for linear reduced problems of the form (2.3.53). Observe that we cannot expect that the reduced computational domain $\Omega^\circ \subset \Omega$,

$$\Omega^\circ = \bigcup_{p \in \mathcal{N}_j^\circ(u_j)} \text{int supp } \lambda_p^{(j)}, \tag{3.38}$$

is resolved by the initial grid $\mathcal{T}_0$. This causes some difficulties concerning the construction of stable splittings of the solution space $\mathcal{S}_j^\circ$. In particular, we have to modify the usual interpolation operator to obtain a counterpart of the well–known hierarchical splitting due to Yserentant [123, 125].

Using the stability of a modified interpolation operator (cf. Hoppe and Kornhuber [75], Kornhuber and Yserentant [88]) and basic results on successive subspace correction methods (cf. Griebel and Oswald [64], Xu [122], and Yserentant [126]), we are able to show that the asymptotic convergence rates of monotone multigrid methods deteriorate at most logarithmically for decreasing meshsize. We emphasize that *no additional regularity assumptions* on the (free) boundary of the reduced domain Ω° enter our considerations. On the other hand, the results are restricted to two space dimensions.

Stable decompositions in arbitrary space dimensions can be obtained by L_2–like projections (cf. Xu [122], Yserentant [126]). However, the stability then relies on certain regularity properties of Ω°. For example, it is sufficient to assume that the "critical region" $\Omega \setminus \Omega^\circ$ is large enough in a certain sense. We will not discuss this subject here, but refer to Kornhuber and Yserentant [88] and Oswald [103] for further information.

3.3.1 A Modified Hierarchical Splitting

For a given subset $\mathcal{N}_j^\bullet(u_j) \subset \mathcal{N}_j$, a nested sequence of reduced subspaces

$$\mathcal{S}_0^\circ \subset \ldots \subset \mathcal{S}_j^\circ$$

is given by

$$\mathcal{S}_k^\circ = \{v \in \mathcal{S}_k \mid v(p) = 0,\ \forall p \in \mathcal{N}_j^\bullet(u_j)\}, \quad k = 0, \ldots, j.$$

To derive a hierarchical splitting of the solution space $\mathcal{S}_j^\circ$ in subspaces $\mathcal{V}_k \subset \mathcal{S}_k^\circ$, $k = 0, \ldots, j$, we will make use of the family of *modified interpolation operators* $I_k : \mathcal{S}_j \to \mathcal{S}_k^\circ$, defined by

$$(I_k v)(p) = \begin{cases} v(p), & \text{if } \lambda_p^{(k)} \in \mathcal{S}_k^\circ \\ 0, & \text{otherwise} \end{cases}, \quad p \in \mathcal{N}_k. \tag{3.39}$$

For all $v \in \mathcal{S}_j^\circ$ we have the decomposition

$$v = I_0 v + \sum_{k=1}^{j} (I_k v - I_{k-1} v) \tag{3.40}$$

so that the space $\mathcal{S}_j^\circ$ is the direct sum of $\mathcal{V}_0 = \mathcal{S}_0^\circ$ and of the subspaces $\mathcal{V}_k$,

$$\mathcal{V}_k = \{I_k v - I_{k-1} v \,|\, v \in \mathcal{S}_j^\circ\} \subset \mathcal{S}_k^\circ. \tag{3.41}$$

We will make use of the H^1–norm $\|v\|_1 = \|v\|_{H^1(\Omega^\circ)}$, given by

$$\|v\|_1 = \left(|v|_1^2 + \|v\|_0^2\right)^{1/2}$$

where $|\cdot|_1$ and $\|\cdot\|_0$ denote the H^1–seminorm and the L^2–norm on Ω°, respectively, and we frequently write

$$\|v\|_{1,t} = \|v\|_{H^1(t)}, \qquad |v|_{1,t} = |v|_{H^1(t)}$$

for the (semi)norms on a triangle t. If not otherwise stated, the various constants $c, c_1, \ldots$, depend only on the initial triangulation $\mathcal{T}_0$ and may have different values at different occurrences.

The following lemma is a variant of the famous discrete L^∞–embedding of Yserentant [123].

Lemma 3.5 *There exists a constant C depending only on the shape regularity of the triangles $t \in \mathcal{T}_k$ such that*

$$|v(x) - v(y)| \le C\sqrt{j-k+1}\,|v|_{1,t} \tag{3.42}$$

holds for all functions $v \in \mathcal{S}_j$ and all points $x, y \in t$.

Proof. We consider the situation as transformed to the usual reference triangle T. From Lemma 2.1 of Yserentant [123], we have

$$|v(x)| \le c\sqrt{j-k+1}\,\|v\|_{1,T}, \quad \forall x \in T,$$

and Poincaré's inequality takes the form

$$\|v\|_{1,T}^2 \le \tfrac{1}{2\pi^2}|v|_{1,T}^2 + 2\,|\int_T v(x)\,dx|.$$

Using the above estimates, we obtain

$$|v(x) - v(y)| = |w(x) - w(y)| \le c\sqrt{j-k+1}\,|v|_{1,T}, \quad \forall x, y \in T,$$

where $w = v - \alpha$ has vanishing mean value on T. Now the assertion follows from the shape regularity and the scaling properties of the H^1–seminorm in two dimensions. □

We emphasize that the estimate (3.42) and following stability estimates are restricted to two space dimensions.

Lemma 3.6 *There exists a constant C depending only on the shape regularity of the triangles in $\mathcal{T}_0$ such that*

$$|I_k v|_1 \le C\sqrt{j-k+1}\,|v|_1 \tag{3.43}$$

holds for all functions $v \in \mathcal{S}_j^\circ$.

Proof. We estimate $|I_k v|^2_{1,t}$ for a fixed $v \in \mathcal{S}_k^\circ$ and all triangles $t \in \mathcal{T}_k$. Two cases have to be distinguished.

In the first case, we assume that t is an "interior" triangle of Ω°, i.e. that the basis functions $\lambda_p^{(k)}$ associated with all three vertices p of t belong to $\mathcal{S}_k^\circ$. Then the restriction of $I_k v$ to t is simply the linear interpolant of v at the vertices of t. Therefore, the estimate

$$|I_k v|_{1,t} \le c_1 \sqrt{j-k+1}\, |v|_{1,t} \tag{3.44}$$

follows from Lemma 3.6 by a simple scaling argument.

If there is a vertex p of t such that $\lambda_p^{(k)} \notin \mathcal{S}_k^\circ$, then the situation is slightly more complicated. In this case, there exists at least one triangle $\bar{t}$ with vertex p that contains a point $\bar{x} \notin \Omega^\circ$. Using $v(\bar{x}) = 0$, we get

$$\begin{aligned} |v(x)| &\le |v(x) - v(p)| + |v(p) - v(\bar{x})| \\ &\le c\sqrt{j-k+1}\ \left(|v|_{1,t} + |v|_{1,\bar{t}} \right) \end{aligned}$$

for all $x \in t$. In a similar way as above, this yields the estimate

$$|I_k v|_{1,t} \le c_2 \sqrt{j-k+1}\ \left(|v|_{1,t} + |v|_{1,\bar{t}} \right). \tag{3.45}$$

As each triangle in $\mathcal{T}_k$ intersects only a finite number of other triangles in $\mathcal{T}_k$, the assertion follows from (3.44) and (3.45). □

The functions v_k in the space $\mathcal{V}_k$ satisfy the estimate

$$4^k \|v_k\|_0^2 \le c\, |v_k|_1^2 \tag{3.46}$$

with a constant c depending again only on the shape regularity of the triangles under consideration. This estimate relies on the observation that every node $p \in \mathcal{N}_k$ has a neighbor $q \in \mathcal{N}_k$ of first or second degree at which the functions in $\mathcal{V}_k$ vanish.

The following stability result is an immediate consequence of Lemma 3.6 and of (3.46) (cf. Yserentant [123, 124]).

Proposition 3.7 *There exists a constant C depending only on the initial triangulation $\mathcal{T}_0$ such that*

$$|I_0 v|_1^2 + \sum_{k=1}^{j} 4^k \|I_k v - I_{k-1} v\|_0^2 \leq C(j+1)^2 |v|_1^2 \tag{3.47}$$

holds for all functions $v \in \mathcal{S}_j^\circ$.

We emphasize that the constant C appearing in (3.47) does not depend on the reduced domain Ω°. Observe that in our application the boundary of Ω° reflects the discrete free boundary separating the different phases and the critical nodes.

3.3.2 Final Convergence Results for the Standard Version

The estimate of the asymptotic convergence rates of the standard monotone multigrid method induced by Λ relies on the decomposition

$$\mathcal{S}_j^\circ = \sum_{\lambda \in \Lambda^\circ} V_\lambda, \quad V_\lambda = \text{span } \{\lambda\}, \tag{3.48}$$

of the solution space in the one–dimensional subspaces V_λ. The following lemma is a corollary of Proposition 3.7.

Lemma 3.8 *For every $v \in \mathcal{S}_j^\circ$ there is a decomposition*

$$v = \sum_{\lambda \in \Lambda^\circ} v_\lambda, \quad v_\lambda \in V_\lambda, \tag{3.49}$$

with the property

$$\sum_{\lambda \in \Lambda^\circ} a_{u_j}(v_\lambda, v_\lambda) \leq c(j+1)^2 a_{u_j}(v, v). \tag{3.50}$$

The constant c depends only on the ellipticity of $a(\cdot,\cdot)$, on the maximal coefficient b_i, $i = 0, \ldots, N$, of Φ (cf. condition (V3)' in Section 1.2.1, p. 23) and on the initial triangulation $\mathcal{T}_0$.

Proof. Consider some fixed $v \in \mathcal{S}_j^\circ$ and let $v = \sum_{k=0}^j v_k$ be the hierarchical splitting (3.40) with $\mathcal{V}_k$ defined in (3.41). Let $\Lambda_{\mathcal{V}_k} \subset \Lambda_k \cap \Lambda^\circ$ be the nodal basis of $\mathcal{V}_k$, $k = 0, \ldots, j$. Using the interpolation $I_{\mathcal{V}_\lambda} : \mathcal{S}_k \to V_\lambda = \operatorname{span}\{\lambda\}$, we define

$$v_\lambda = I_{V_\lambda} v_k, \quad \lambda \in \Lambda_{\mathcal{V}_k}, \;\; k = 0, \ldots, j,$$

and $v_\lambda = 0$ for all remaining $\lambda \in \Lambda^\circ$. This leads to the decomposition

$$v = \sum_{k=0}^{j} v_k = \sum_{k=0}^{j} \sum_{\lambda \in \Lambda_{\mathcal{V}_k}} v_\lambda = \sum_{\lambda \in \Lambda^\circ} v_\lambda. \tag{3.51}$$

We will show that the splitting (3.51) satisfies (3.50). To this end, we use Proposition 3.7, the estimate

$$\tfrac{1}{6} \sum_{t \in \mathcal{T}_k} |t| \sum_{p \in t} |v_k(p)|^2 \le 2 \, \|v_k\|_0^2,$$

the well–known inverse inequality

$$|v_\lambda|_1^2 \le c 4^k \|v_\lambda\|_0^2,$$

and the equivalence of norms on $\mathcal{S}_0$, to obtain

$$\sum_{\lambda \in \Lambda^\circ} |v_\lambda|_1^2 \le c(j+1)^2 |v|_1^2. \tag{3.52}$$

Now the assertion follows from the equivalence

$$c \, a_{u_j}(v, v) \le |v|_1^2 \le C a_{u_j}(v, v), \tag{3.53}$$

with constants c, C depending only on the ellipticity of $a(\cdot, \cdot)$ and on the maximal coefficient b_i, $i = 0, \ldots, N$, of Φ. □

The following lemma is an immediate consequence of a strengthened Cauchy-Schwarz inequality (see e.g. Bornemann [24] or Zhang [127]).

Lemma 3.9 *Assume that $v \in \mathcal{S}_j$ is decomposed according to*

$$v = \sum_{k=0}^{j} v_k, \quad v_k \in \textit{span}\ \Lambda^{(k)},$$

denoting $\Lambda^{(0)} = \Lambda_0$ and $\Lambda^{(k)} = \Lambda_k \setminus \Lambda_{k-1}$, $k = 1, \ldots, j$. Then the estimate

$$|v|_1^2 \le C \sum_{k=0}^{j} |v_k|_1^2 \tag{3.54}$$

holds with a constant C depending only on the initial triangulation $\mathcal{T}_0$.

Recall that the discrete phases of some $v \in \mathcal{S}_j$ are defined in Section 2.3 where also the non–degeneracy condition (2.3.42) is stated. The piecewise quadratic function Φ generates the functional ϕ_j (cf. (1.3.55)), and we assume that Φ satisfies the conditions (V1), (V2), and (V3)' from Section 1.2.1), p. 23. Now we are ready to state the final convergence theorem for our standard multigrid method.

Theorem 3.10 *The standard monotone multigrid method induced by Λ is globally convergent.*

Assume that the discrete problem (1.3.57) is non–degenerate. Then the discrete phases of the iterates $(u_j^\nu)_{\nu \ge 0}$ converge to the discrete phases of u_j, and there is a $\nu_0 = \nu_0(j) \ge 0$ such that we have the error estimate

$$|||u_j - u_j^{\nu+1}||| \le (1 - c(j+1)^{-4})|||u_j - u_j^\nu|||, \quad \forall \nu \ge \nu_0, \tag{3.55}$$

with respect to the asymptotic energy norm $|||\cdot||| = a_{u_j}(\cdot,\cdot)^{1/2}$. The positive constant $c < 1$ depends only on the ellipticity of $a(\cdot,\cdot)$, on the maximal coefficient b_i, $i = 0, \ldots, N$, from (V3)' and on the initial triangulation $\mathcal{T}_0$.

Proof. The global convergence follows from Proposition 3.3 and, in the non–degenerate case, the iteration reduces asymptotically to the successive subspace correction method characterized by the splitting (3.48).

Note that the number $J = \#\Lambda^\circ$ of one–dimensional subspaces V_λ satisfies

$$J \le c\, 4^j. \tag{3.56}$$

Following Griebel and Oswald [64], we introduce the norm $\|\!| \cdot |\!|$,

$$|\!| v |\!|^2 = \inf_{\{v_\lambda \in V_\lambda:\ v = \sum_{\lambda \in \Lambda^\circ} v_\lambda\}} \sum_{\lambda \in \Lambda^\circ} a_{u_j}(v_\lambda, v_\lambda).$$

As a consequence of Lemma 3.8, we have the *lower estimate*

$$|\!| v |\!|^2 \le c(j+1)^2 a_{u_j}(v,v), \quad \forall v \in \mathcal{S}_j^\circ, \tag{3.57}$$

with a suitable constant c. Using Lemma 3.9, the equivalence (3.53) and a simple coloring argument (cf. e.g. Bornemann [24] or Zhang [127]), we obtain the converse *upper estimate*

$$a_{u_j}(v,v) \le C\, |\!| v |\!|^2, \quad \forall v \in \mathcal{S}_j^\circ. \tag{3.58}$$

By virtue of (3.56), (3.57) and (3.58), the assertion follows from Theorem 4 by Griebel and Oswald [64]. □

We emphasize that the estimate (3.55) describes the worst case. Absolutely no regularity assumptions on the continuous or discrete free boundary enter the constant c. In addition, we have considered the most simple variant of standard monotone multigrid methods.

By repeating the successive (approximate) minimization in the direction of the basis functions $\lambda_p^{(k)}$ on each level $k = j, \cdots, 0$ in reversed order, we obtain a *standard monotone multigrid method with symmetric smoother*. For this variant, we get a $\mathcal{O}(j^2(\log j)^2)$ estimate.

Together with the stability (3.43) of the modified interpolation operators I_k, all these results are restricted to two space dimensions. In the higher dimensional case, we have to work with modified L^2-projections. In contrast to the interpolation technique, this requires a certain regularity of the (free) boundary of the reduced computational domain Ω°. For example, excluding "poor" critical regions $\Omega \setminus \Omega^\circ$, such as lines or points, Kornhuber and Yserentant [88] proved an $\mathcal{O}(j)$ estimate. Mesh–independent convergence rates in arbitrary space dimensions were stated by Oswald [103] under very similar conditions.

Upper bounds for the *global convergence rates* which deteriorate exponentially with the number of refinement levels j can be obtained from the work of Tai [115]. However, these results are much too pessimistic as compared to our numerical experiments where we usually found *uniform* upper bounds for the global convergence rates. A theoretical justification of these observations is still an open problem.

3.3.3 Final Convergence Results for the Truncated Version

We state an analogue of Theorem 3.10 for the corresponding truncated multigrid method.

Theorem 3.11 *The truncated monotone multigrid method induced by* $(\tilde{\Lambda}^\nu)_{\nu\geq 0}$ *is globally convergent.*

Assume that the discrete problem (1.3.57) is non–degenerate. Then the discrete phases of the iterates $(u_j^\nu)_{\nu\geq 0}$ *converge to the discrete phases of* u_j*, and there is a* $\nu_0 = \nu_0(j) \geq 0$ *such that we have the error estimate*

$$|||u_j - u_j^{\nu+1}||| \leq (1 - c(j+1)^{-6})|||u_j - u_j^\nu|||, \quad \forall \nu \geq \nu_0, \tag{3.59}$$

with respect to the asymptotic energy norm $|||\cdot||| = a_{u_j}(\cdot,\cdot)^{1/2}$*. The positive constant* $c < 1$ *depends only on the ellipticity of* $a(\cdot,\cdot)$*, on the maximal coefficient* b_i*,* $i = 0,\ldots,N$*, from (V3)' and on the initial triangulation* $\mathcal{T}_0$*.*

Proof. We only have to show the asymptotic error estimate (3.59). By virtue of Proposition 3.4, the iteration reduces asymptotically to the successive subspace correction method characterized by the splitting

$$\mathcal{S}_j^\circ = \sum_{\lambda\in\tilde{\Lambda}^\circ} V_\lambda, \quad v_\lambda = \text{span}\,\{\lambda\}. \tag{3.60}$$

In analogy to the standard case, this splitting gives rise to the norm

$$\|\, v \,\|^2 = \inf_{\{v_\lambda\in V_\lambda:\ v=\sum_{\lambda\in\tilde{\Lambda}^\circ} v_\lambda\}} \sum_{\lambda\in\tilde{\Lambda}^\circ} a_{u_j}(v_\lambda, v_\lambda).$$

In order to apply the results of Griebel and Oswald [64], we again have to provide a lower and an upper estimate of the form (3.57) and (3.58), respectively. As a consequence of $\Lambda^\circ \subset \tilde{\Lambda}^\circ$, the lower estimate is trivial.

For the upper estimate, we cannot apply Lemma 3.9, because truncations of the basis functions $\lambda_p^{(k)} \in \Lambda_k$ are in general not contained in $\mathcal{S}_k$. Using the triangle inequality and the Cauchy–Schwarz inequality, we can show that

$$|v|_1^2 \leq c\,(j+1)\,\sum_{k=0}^{j} |\tilde{v}_k|_1^2 \tag{3.61}$$

holds for all decompositions $v = \tilde{v}_0 + \ldots + \tilde{v}_j$ of $v \in \mathcal{S}_j$ in fine grid functions $\tilde{v}_k \in \mathcal{S}_j$, $k = 0, \ldots, j$. Replacing (3.54) by this more pessimistic estimate, we get the upper estimate

$$a_{u_j}(v, v) \leq c(j+1) \parallel v \parallel^2, \quad \forall v \in \mathcal{S}_j. \tag{3.62}$$

Now the assertion (3.59) follows again from Theorem 4 in [64]. □

Again the pessimistic bound (3.59) can be improved for more sophisticated variants. In analogy to the standard case, we can define a *truncated monotone multigrid method with symmetric smoother*. For this method, we can prove an $\mathcal{O}(j^3)$ estimate in a similar way as above (see e.g. Theorem 4.5).

Nevertheless, these asymptotic estimates seem to contradict our intention to improve the convergence rates by suitable truncation. The reason is that the present theory is not elaborated enough to work out the possible benefits of such unconventional extensions for the lower estimate. On the other hand, the resulting problems for the upper estimate are felt immediately.

In several numerical experiments, we found a very similar asymptotic behavior *with respect to* j both of standard and truncated methods. This was usually better than predicted by (3.55) and (3.59), respectively. However, the convergence of truncated methods was always much faster than the convergence of the corresponding standard variant. For uniformly refined triangulations, we refer to the recent papers of Kornhuber [82, 83] and Kornhuber and Yserentant [88]. Non–uniform refinements will be considered in Chapter 5.

4 A Posteriori Error Estimates and Adaptive Refinement

In many practical computations, we want to approximate the solution u of the continuous minimization problem (1.2.33) up to a prescribed tolerance TOL. In this case, the discretization (1.3.57) has to be chosen in such a way that we can compute an approximation $\tilde{u}_J \in \mathcal{S}_J$ of the exact finite element solution $u_J \in \mathcal{S}_J$ with the property

$$\|u - \tilde{u}_J\| \leq \text{TOL}.$$

Adaptive methods are intended to produce a triangulation $\mathcal{T}_J$ which provides the desired accuracy with a minimal number of nodes. This involves the (approximate) solution of a nonlinear approximation problem (cf. Oswald [104]). Adaptive multilevel methods, to be considered in this chapter, provide a well–established framework for such algorithms which can be briefly described as follows.

Assume that an intentionally coarse initial triangulation $\mathcal{T}_0$ is available. Though the actual construction of such a $\mathcal{T}_0$ may be a non–trivial task, we will not discuss this problem here (see for example Kornhuber and Roitzsch [87, 106]).

Starting with $\mathcal{T}_0$, a sequence of successively refined triangulations is produced in the following way. We discretize the continuous problem with respect to the actual triangulation $\mathcal{T}_j$ and solve the resulting discrete problem up to a certain accuracy, to obtain the actual approximation $\tilde{u}_j$. If the global accuracy condition $\|u - \tilde{u}_j\| \leq \text{TOL}$ is satisfied, then the approximation $\tilde{u}_j$ is accepted and the computation is stopped. Otherwise we adapt $\mathcal{T}_j$ by local refinement of suitably selected triangles and repeat the same cycle on the next refinement level.

In this way, we simultaneously get an increasing sequence of *nested triangulations* $\mathcal{T}_0, \ldots, \mathcal{T}_j$, $j \leq J$, which can be directly used for the fast solution of the corresponding discrete problems by monotone multigrid methods (cf. Chapter 3). This is not the case for related strategies based on a complete

remeshing of the computational domain (see for example Nochetto, Paolini, and Verdi [98, 99]). Moreover, successive local refinement extends without difficulties to three (or more) space dimensions (cf. e.g. Bänsch [13], Bey [22], Bornemann, Erdmann, and Kornhuber [27]). In this case, remeshing becomes considerably more time–consuming.

Using the general approach described above, we will derive a posteriori estimates for the *approximation error* $\|u - \tilde{u}_j\|$, a refinement strategy based on local error indicators and stopping criteria for the complete adaptive algorithm. As a discrete counterpart, we will provide a posteriori estimates for the *algebraic error* $\|u_j - \tilde{u}_j\|$ together with stopping criteria for the iterative solution on each refinement level.

4.1 A Posteriori Estimates of the Approximation Error

Let $u \in H = H_0^1(\Omega)$ denote the exact solution of the continuous minimization problem (1.2.33), $u_j \in \mathcal{S}_j$ the exact solution of the approximate discrete problem (1.3.57) and $\tilde{u}_j \in \mathcal{S}_j$ an approximate solution of (1.3.57). Usually, $\tilde{u}_j$ results from a certain number of steps of some iterative solver. As only $\tilde{u}_j$ is available in actual computations, we will concentrate on reliable and efficient a posteriori estimates providing upper and lower bounds of the *approximation error* $\|u - \tilde{u}_j\|$. Related estimates of the *algebraic error* $\|u_j - \tilde{u}_j\|$ will be derived in the final section of this chapter.

The construction of error estimates will be carried out in two steps.

- Discretize the defect problem with respect to an enlarged space.
- Localize the resulting discrete defect problem.

In the case of elliptic selfadjoint problems, the resulting *hierarchic error estimates* were introduced by Zienkiewicz, Gago and Kelly [128]. The intimate relation of error estimation and preconditioning first appeared explicitly in a paper of Deuflhard, Leinen and Yserentant [45]. Replacing the piecewise quadratic functions by other extensions of the original space $\mathcal{S}_j$, the hierarchical approach allows a unified view on a variety of apparently different concepts (cf. Bornemann, Erdmann and Kornhuber [28] and Verfürth [118]). A slightly modified extension is of particular interest in the case of three

space dimensions [28]. For a comprehensive overview on other approaches to a posteriori error estimation, we refer to the monograph of Verfürth [119] in this series.

Using Newton type linearization, Bank and Smith [10] have extended hierarchical error estimates from the elliptic selfadjoint case to differentiable nonlinear problems. As we cannot use Newton's method for our non–smooth minimization problem, we will apply the hierarchical concept directly. This requires some care in the localization of the discrete defect problem. A straightforward approach was applied successfully to a special obstacle problem arising from semiconductor device simulation (cf. Kornhuber and Roitzsch [86, 87]). However, it turned out in the subsequent analysis that the resulting error estimator is not robust (cf. Hoppe and Kornhuber [75]). In particular, the localization step may produce a vanishing error estimate even though the solution of the original discrete defect problem is not zero.

It will turn out that this problem can be remedied by a *diagonal scaling* of the discrete defect problem. In this way, the original global problem is decomposed in a number of one–dimensional subproblems. The constants describing the quality of the error estimates are independent of the refinement level j, if the discrete error is a *high–frequency function*. Recall similar properties of cascadic iterations (cf. Deuflhard [44], Shaidurov [112] and Bornemann and Deuflhard [26]). This heuristic assumption can be justified under severe restrictions on the discrete defect problem. There is numerical evidence indicating that these conditions may be relaxed. We refer to Kornhuber [84, 85] and to the numerical results presented in the final chapter.

A more robust (and unfortunately more expensive) a posteriori error estimate is based on *nonlinear iteration*. In this case we get uniform constants under (slightly) less restrictive conditions. An extension of these results to more realistic situations is very closely related to optimal global bounds for the convergence rates of monotone multigrid methods. Recall that this is an open problem.

4.1.1 A Discrete Defect Problem

For a given $\tilde{u}_j \in \mathcal{S}_j$ the desired correction $e = u - \tilde{u}_j \in H$ is the solution of the *defect problem*

$$\begin{aligned} e \in H: \quad & \tfrac{1}{2}a(e,e) - r(e) + \psi(e) \\ & \leq \tfrac{1}{2}a(v,v) - r(v) + \psi(v), \quad \forall v \in H, \end{aligned} \tag{1.1}$$

where we have used the translated functional $\psi : H \to \mathbb{R} \cup \{+\infty\}$,

$$\psi(v) = \phi(\tilde{u}_j + v) = \int_\Omega \Phi(\tilde{u}_j + v)\, dx,$$

and the bounded linear functional $r : H \to \mathbb{R}$,

$$r(v) = \ell(v) - a(\tilde{u}_j, v).$$

In order to discretize the continuous defect problem (1.1), we introduce the space $\mathcal{Q}_j \subset H$ of continuous, piecewise quadratic functions, spanned by the nodal basis

$$\Lambda_j^{\mathcal{Q}} = \{\lambda_p^{\mathcal{Q}} \mid p \in \mathcal{N}_{\mathcal{Q}}\},$$

where $\mathcal{N}_{\mathcal{Q}} = \mathcal{N}_j \cup \mathcal{N}_{\mathcal{E}}$ and $\mathcal{N}_{\mathcal{E}}$ consists of the midpoints of the interior edges $e \in \mathcal{E}_j$. Interpolating $\Phi(\tilde{u}_j + v)$ by piecewise quadratic finite elements, we obtain the approximation $\psi_{\mathcal{Q}} : \mathcal{Q}_j \to \mathbb{R} \cup \{+\infty\}$,

$$\psi_{\mathcal{Q}}(v) = \sum_{p \in \mathcal{N}_{\mathcal{Q}}} \Phi(\tilde{u}_j(p) + v(p)) \int_\Omega \lambda_p^{\mathcal{Q}}(x) dx,$$

of the functional ψ. Then $e_{\mathcal{Q}}$ is the unique solution of the *discrete defect problem*

$$\begin{aligned} e_{\mathcal{Q}} \in \mathcal{Q}_j : \quad & \tfrac{1}{2} a(e_{\mathcal{Q}}, e_{\mathcal{Q}}) - r(e_{\mathcal{Q}}) + \psi_{\mathcal{Q}}(e_{\mathcal{Q}}) \\ & \qquad \le \tfrac{1}{2} a(v, v) - r(v) + \psi_{\mathcal{Q}}(v), \quad \forall v \in \mathcal{Q}_j. \end{aligned} \tag{1.2}$$

In the light of Theorem 1.6 and Proposition 1.7, the minimization problem (1.2) can be reformulated as the variational inequality

$$\begin{aligned} e_{\mathcal{Q}} \in \mathcal{Q}_j : \quad & a(e_{\mathcal{Q}}, v - e_{\mathcal{Q}}) + \psi_{\mathcal{Q}}(v) - \psi_{\mathcal{Q}}(e_{\mathcal{Q}}) \\ & \qquad \ge r(v - e_{\mathcal{Q}}), \quad \forall v \in \mathcal{Q}_j, \end{aligned} \tag{1.3}$$

or as the variational inclusion

$$e_{\mathcal{Q}} \in \mathcal{Q}_j : \quad 0 \in a(e_{\mathcal{Q}}, v) - r(v) + \partial \psi_{\mathcal{Q}}(e_{\mathcal{Q}})(v), \quad \forall v \in \mathcal{Q}_j. \tag{1.4}$$

Using Green's formula, the local residuals $r(\lambda_p^{\mathcal{Q}})$, $p \in \mathcal{N}_{\mathcal{Q}}$, can be reformulated in terms of *local consistency errors* and *jumps of the normal fluxes*

of $\tilde{u}_j$. This representation is frequently used in other residual based error estimates (cf. e.g. Verfürth [119]).

Correcting $\tilde{u}_j$ by $e_{\mathcal{Q}}$, we obtain the *piecewise quadratic approximation*

$$u_{\mathcal{Q}} = \tilde{u}_j + e_{\mathcal{Q}} \in \mathcal{Q}_j$$

which is the unique solution of the minimization problem

$$u_{\mathcal{Q}} \in \mathcal{Q}_j : \quad \mathcal{J}(u_{\mathcal{Q}}) + \phi_{\mathcal{Q}}(u_{\mathcal{Q}}) \le \mathcal{J}(v) + \phi_{\mathcal{Q}}(v), \quad \forall v \in \mathcal{Q}_j, \tag{1.5}$$

where the approximation $\phi_{\mathcal{Q}} : \mathcal{Q}_j \to \mathbb{R} \cup \{\infty\}$ of ϕ is defined by

$$\phi_{\mathcal{Q}}(\tilde{u}_j + v) = \psi_{\mathcal{Q}}(v).$$

The condition

$$\text{(S)} \qquad \|u - u_{\mathcal{Q}}\| \le \beta \|u - \tilde{u}_j\|, \quad \beta < 1,$$

states that (1.5) provides a *better approximation* $u_{\mathcal{Q}} \in \mathcal{Q}_j$ of u than the given function $\tilde{u}_j \in \mathcal{S}_j$.

Proposition 4.1 *Assume that the condition (S) is satisfied and let $e_{\mathcal{Q}}$ be the solution of the discrete defect problem (1.2). Then we have the estimates*

$$(1+\beta)^{-1}\|e_{\mathcal{Q}}\| \le \|u - \tilde{u}_j\| \le (1-\beta)^{-1}\|e_{\mathcal{Q}}\|. \tag{1.6}$$

Proof. We only show the lower bound for $\|u - \tilde{u}_j\|$ which immediately follows from (S) and the triangle inequality

$$\|u - \tilde{u}_j\| \ge \|u_{\mathcal{Q}} - \tilde{u}_j\| - \|u - u_{\mathcal{Q}}\|.$$

□

The crucial condition (S) with $\beta = \beta_s/(1-\beta_a) < 1$ is a consequence of the *saturation assumption*

$$\|u - u_{\mathcal{Q}}\| \le \beta_s \|u - u_j\|, \quad \beta_s < 1, \tag{1.7}$$

and of the *algebraic accuracy assumption*

$$\|u_j - \tilde{u}_j\| \le \beta_a \|u - u_j\|, \quad \beta_a < 1 - \beta_s. \tag{1.8}$$

The saturation assumption (1.7) states that the larger finite element space $\mathcal{Q}_j$ provides a better approximation than the original space $\mathcal{S}_j$. For sufficiently regular problems, the piecewise quadratic solution $u_{\mathcal{Q}}$ is even an approximation of higher order (cf. e.g. Brezzi, Hager, and Raviart [36]). In this case, (1.7) clearly holds for sufficiently fine triangulations. On the other hand, there are simple examples showing that (1.7) may be violated, if the mesh is not properly chosen. In this sense, reliable a posteriori error estimates still involve a certain amount of *a priori information*.

The algorithmic realization of the algebraic accuracy assumption (1.8) will be discussed in the final chapter.

In the case of elliptic selfadjoint problems, (1.8) is not needed and the saturation assumption (1.7) is even equivalent to the upper estimate in (1.6) with $\beta = \beta_s$. We refer to Bornemann, Erdmann and Kornhuber [28] for details.

4.1.2 An Error Estimate Based on Preconditioning

In general, the solution of the discrete defect problem (1.2) is not available at reasonable computational cost. Hence, we consider further approximations of (1.2) which should preserve lower and upper bounds of the form (1.6).

Extending well–known results from the elliptic selfadjoint case (cf. Deuflhard, Leinen and Yserentant [45], Bornemann, Erdmann and Kornhuber [28], and Bank and Smith [10]), we will now investigate the *effect of preconditioning* on the solution $e_{\mathcal{Q}}$ of (1.3). For this reason, we consider the variational inequality

$$\begin{aligned} \tilde{e}_{\mathcal{Q}} \in \mathcal{Q}_j : \quad \tilde{a}(\tilde{e}_{\mathcal{Q}}, v - \tilde{e}_{\mathcal{Q}}) + \psi_{\mathcal{Q}}(v) - \psi_{\mathcal{Q}}(\tilde{e}_{\mathcal{Q}}) \\ \geq r(v - \tilde{e}_{\mathcal{Q}}), \quad \forall v \in \mathcal{Q}_j, \end{aligned} \tag{1.9}$$

with some symmetric and positive definite bilinear form $\tilde{a}(\cdot,\cdot)$ on $\mathcal{Q}_j$. Observe that the *preconditioned defect problem* (1.9) is uniquely solvable and that the preconditioner $\tilde{a}(\cdot,\cdot)$ induces the norm $|\cdot| = \tilde{a}(\cdot,\cdot)^{1/2}$ on $\mathcal{Q}_j$.

Proposition 4.2 *Assume that the norm equivalence*

$$\gamma_0 \tilde{a}(v,v) \leq a(v,v) \leq \gamma_1 \tilde{a}(v,v), \quad \forall v \in \mathit{span}\{e_{\mathcal{Q}}, \tilde{e}_{\mathcal{Q}}\}, \tag{1.10}$$

holds with positive constants γ_0, γ_1. *Then we have the estimates*

$$c_0|\tilde{e}_{\mathcal{Q}}|^2 \le \|e_{\mathcal{Q}}\|^2 \le c_1|\tilde{e}_{\mathcal{Q}}|^2 \tag{1.11}$$

with $c_0 = (\gamma_0^{-1} + 2\gamma_1(1+\gamma_0^{-1}))^{-1}$ *and* $c_1 = \gamma_1 + 2\gamma_0^{-1}(1+\gamma_1)$.

Proof. By symmetry arguments it is sufficient to establish only the right inequality in (1.11). Inserting $v = \tilde{e}_{\mathcal{Q}}$ in the original discrete defect problem (1.3), we obtain

$$\|e_{\mathcal{Q}}\|^2 \le a(e_{\mathcal{Q}}, \tilde{e}_{\mathcal{Q}}) + \psi_{\mathcal{Q}}(\tilde{e}_{\mathcal{Q}}) - \psi_{\mathcal{Q}}(e_{\mathcal{Q}}) + r(e_{\mathcal{Q}} - \tilde{e}_{\mathcal{Q}}).$$

Now the inequality $2a(e_{\mathcal{Q}}, \tilde{e}_{\mathcal{Q}}) \le \|e_{\mathcal{Q}}\|^2 + \|\tilde{e}_{\mathcal{Q}}\|^2$ and (1.10) yield

$$\|e_{\mathcal{Q}}\|^2 \le \gamma_1|\tilde{e}_{\mathcal{Q}}|^2 + 2(\psi_{\mathcal{Q}}(\tilde{e}_{\mathcal{Q}}) - \psi_{\mathcal{Q}}(e_{\mathcal{Q}}) + r(e_{\mathcal{Q}} - \tilde{e}_{\mathcal{Q}})). \tag{1.12}$$

It remains to show that

$$\psi_{\mathcal{Q}}(\tilde{e}_{\mathcal{Q}}) - \psi_{\mathcal{Q}}(e_{\mathcal{Q}}) + r(e_{\mathcal{Q}} - \tilde{e}_{\mathcal{Q}}) \le \gamma_0^{-1}(\gamma_1 + 1)|\tilde{e}_{\mathcal{Q}}|^2. \tag{1.13}$$

Inserting $v = e_{\mathcal{Q}}$ in (1.9) and using the Cauchy–Schwarz inequality, we get

$$\psi_{\mathcal{Q}}(\tilde{e}_{\mathcal{Q}}) - \psi_{\mathcal{Q}}(e_{\mathcal{Q}}) + r(e_{\mathcal{Q}} - \tilde{e}_{\mathcal{Q}}) \le |\tilde{e}_{\mathcal{Q}}||e_{\mathcal{Q}} - \tilde{e}_{\mathcal{Q}}|$$

so that (1.13) follows from

$$|e_{\mathcal{Q}} - \tilde{e}_{\mathcal{Q}}| \le \gamma_0^{-1}(1+\gamma_1)|\tilde{e}_{\mathcal{Q}}|. \tag{1.14}$$

In order to prove (1.14), we again insert $v = \tilde{e}_{\mathcal{Q}}$ in (1.3) and $v = e_{\mathcal{Q}}$ in the preconditioned problem (1.9). Adding the two resulting inequalities, we obtain

$$a(e_{\mathcal{Q}}, \tilde{e}_{\mathcal{Q}} - e_{\mathcal{Q}}) + \tilde{a}(\tilde{e}_{\mathcal{Q}}, e_{\mathcal{Q}} - \tilde{e}_{\mathcal{Q}}) \ge 0$$

which can be reformulated as

$$\|\tilde{e}_{\mathcal{Q}} - e_{\mathcal{Q}}\|^2 \le a(\tilde{e}_{\mathcal{Q}}, \tilde{e}_{\mathcal{Q}} - e_{\mathcal{Q}}) - \tilde{a}(\tilde{e}_{\mathcal{Q}}, \tilde{e}_{\mathcal{Q}} - e_{\mathcal{Q}}).$$

The assertion now follows from the Cauchy–Schwarz inequality and (1.10). □

In the light of Proposition 4.2, we are left with the problem of selecting a preconditioner $\tilde{a}(\cdot,\cdot)$ which combines reasonable constants γ_0, γ_1 with a cheap evaluation of $\tilde{e}_{\mathcal{Q}}$. In analogy to the linear selfadjoint case, one might be tempted to construct a preconditioner based on the hierarchical splitting

$$\mathcal{Q} = \mathcal{S} \oplus \mathcal{V} \tag{1.15}$$

where the difference space $\mathcal{V} = \text{span}\{\lambda_p^{\mathcal{Q}} \mid p \in \mathcal{N}_{\mathcal{E}}\}$ consists of the quadratic bubble functions associated with the edges $\mathcal{E}$ (cf. e.g. Deuflhard, Leinen and Yserentant [45], Bornemann, Erdmann and Kornhuber [27], Bank and Smith [10]). However, in contrast to the linear case, the unknowns now become coupled with respect to the functional $\psi_{\mathcal{Q}}$ as soon as the hierarchical representation is used (cf. Hoppe and Kornhuber[75], Erdmann, Frei, Hoppe, Kornhuber and Wiest [52], and Erdmann, Hoppe and Kornhuber [53]). Even in simple cases, this coupling cannot be ignored without losing the reliability of the resulting error estimate (cf. Hoppe and Kornhuber [75]). On the other hand, the coupled preconditioned problem is still not solvable with reasonable computational cost.

To find a way out of this dilemma, observe that the constants γ_0, γ_1 appearing in the crucial estimate (1.11) depend only on the *local quality* of the preconditioner $\tilde{a}(\cdot,\cdot)$ on the subspace span $\{e_{\mathcal{Q}}, \tilde{e}_{\mathcal{Q}}\} \subset \mathcal{Q}_j$. As a consequence, we can expect good results even from very simple preconditioners like the diagonal scaling

$$\tilde{a}(v,w) = \sum_{p \in \mathcal{N}_{\mathcal{Q}}} v(p)w(p)a(\lambda_p^{\mathcal{Q}}, \lambda_p^{\mathcal{Q}}), \tag{1.16}$$

if $e_{\mathcal{Q}}$ and $\tilde{e}_{\mathcal{Q}}$ are *high frequency* functions.

In addition, the preconditioned defect equation (1.9) resulting from the diagonal scaling (1.16) consists of the *independent local subproblems*

$$e_p \in \text{span}\,\{\lambda_p^{\mathcal{Q}}\}: \quad 0 \in \tilde{a}(e_p, \lambda_p^{\mathcal{Q}}) - r(\lambda_p^{\mathcal{Q}}) + \partial\psi_{\mathcal{Q}}(e_p)(\lambda_p^{\mathcal{Q}}), \tag{1.17}$$

for $p \in \mathcal{N}_{\mathcal{Q}}$. Exploiting that Φ is piecewise quadratic in the sense of condition (V3)' (cf. Section 1.2.1, p. 23), we can solve all these sub–problems explicitly. We refer to similar results in Section 2.1.1. Finally, the approximate correction

$$\tilde{e}_{\mathcal{Q}} = \sum_{p \in \mathcal{N}_{\mathcal{Q}}} e_p \tag{1.18}$$

provides the *local error estimate* $|\tilde{e}_{\mathcal{Q}}|$.

From the Propositions 4.1 and 4.2, we immediately get a (quite pessimistic) upper bound for the effectivity index $\kappa = |\tilde{e}_{\mathcal{Q}}|/\|e\|$ of the resulting error estimate which increases exponentially with the refinement level. However, the localization at least preserves a *non–vanishing* error estimate $|\tilde{e}_{\mathcal{Q}}| \neq 0$, if $e_{\mathcal{Q}}$ is not zero. Recall that related previous error estimates do not have this property (cf. Hoppe and Kornhuber [75]).

The heuristic assumption that $\mathrm{span}\{e_{\mathcal{Q}}, \tilde{e}_{\mathcal{Q}}\}$ consists of high–frequency functions can be justified theoretically, if the following conditions are fulfilled.

(P1) $\tilde{u}_j = u_j$

(P2) The discrete phases of u_j, $u_j + \tilde{e}_{\mathcal{Q}}$ and $u_{\mathcal{Q}}$ coincide.

(P3) The subdifferentials $\partial\phi_j(u_j)$, $\partial\phi_{\mathcal{Q}}(u_j)$ coincide on $\mathcal{S}_j^\circ \subset \mathcal{S}_j$.

(P4) The reduced subspace $\mathcal{S}_j^\circ$ is contained in the reduced subspace $\mathcal{Q}_j^\circ$.

Recall the definition of discrete phases in Section 2.3. Here, the discrete phases of $v \in \mathcal{Q}_j$ are given by

$$\mathcal{N}_{\mathcal{Q}}^i(v) = \{p \in \mathcal{N}_{\mathcal{Q}} \mid v(p) \in (\theta_i, \theta_{i+1})\}, \quad i = 0, \ldots, N,$$

and the reduced space $\mathcal{Q}_j^\circ \subset \mathcal{Q}_j$,

$$\mathcal{Q}_j^\circ = \{v \in \mathcal{Q}_j \mid v(p) = 0, \forall p \in \mathcal{N}_{\mathcal{Q}}^\bullet\}, \quad \mathcal{N}_{\mathcal{Q}}^\bullet = \mathcal{N}_{\mathcal{Q}} \setminus \bigcup_{i=0}^{N} \mathcal{N}_{\mathcal{Q}}^i(u_{\mathcal{Q}}),$$

is defined in analogy to $\mathcal{S}_j^\circ \subset \mathcal{S}_j$ (cf. (2.3.54) in Section 2.3).

Theorem 4.3 *Assume that the conditions (P1)–(P4) are satisfied. Then we have the equivalence*

$$c|\tilde{e}_{\mathcal{Q}}| \leq \|e_{\mathcal{Q}}\| \leq C|\tilde{e}_{\mathcal{Q}}|. \tag{1.19}$$

The constants c, C depend only on the ellipticity of $a(\cdot,\cdot)$, on the maximal coefficient b_i, $i = 0, \ldots, N$, from (V3)' on p. 23, and on the initial triangulation $\mathcal{T}_0$.

Proof. As a consequence of (P1) and (P2), the correction $e_{\mathcal{Q}}$ is the unique solution of the variational equality

$$\begin{aligned} e_{\mathcal{Q}} \in \mathcal{Q}_j^\circ : \quad & a(e_{\mathcal{Q}}, v) + b_{u_{\mathcal{Q}}}(e_{\mathcal{Q}}, v) \\ & = r(v) + f_{u_{\mathcal{Q}}}(v) - b_{u_{\mathcal{Q}}}(u_j, v), \quad \forall v \in \mathcal{Q}_j^\circ, \end{aligned} \tag{1.20}$$

where the bilinear form $b_{u_{\mathcal{Q}}}(\cdot,\cdot)$ and the functional $f_{u_{\mathcal{Q}}}$ are defined in analogy to (2.3.50) and (2.3.51). In particular, we have

$$\partial\phi_{\mathcal{Q}}(u_{\mathcal{Q}})(v) = b_{u_{\mathcal{Q}}}(u_{\mathcal{Q}}, v) - f_{u_{\mathcal{Q}}}(v), \quad \forall v \in \mathcal{Q}_j^\circ.$$

Using (P2) and (P3), we also obtain

$$\partial\phi_j(u_j)(v) = b_{u_{\mathcal{Q}}}(u_j, v) - f_{u_{\mathcal{Q}}}(v), \quad \forall v \in \mathcal{S}_j^\circ. \tag{1.21}$$

Hence, being the exact solution of the discrete problem (1.3.57), u_j satisfies the variational equality

$$a(u_j, v) + b_{u_{\mathcal{Q}}}(u_j, v) = \ell(v) + f_{u_{\mathcal{Q}}}(v), \quad \forall v \in \mathcal{S}_j^\circ. \tag{1.22}$$

Exploiting (P2), we can show in a similar way as (1.20) that $\tilde{e}_{\mathcal{Q}}$ is the unique solution of the variational equality

$$\begin{aligned} \tilde{e}_{\mathcal{Q}} \in \mathcal{Q}_j^\circ : \quad & \tilde{a}(\tilde{e}_{\mathcal{Q}}, v) + b_{u_{\mathcal{Q}}}(\tilde{e}_{\mathcal{Q}}, v) \\ & = r(v) + f_{u_{\mathcal{Q}}}(v) - b_{u_{\mathcal{Q}}}(u_j, v), \ \ \forall v \in \mathcal{Q}_j^\circ. \end{aligned} \tag{1.23}$$

Introducing the bilinear forms

$$a_{u_{\mathcal{Q}}}(v, w) = a(v, w) + b_{u_{\mathcal{Q}}}(v, w), \ \ \tilde{a}_{u_{\mathcal{Q}}}(v, w) = \tilde{a}(v, w) + b_{u_{\mathcal{Q}}}(v, w)$$

and the residual

$$r_{u_{\mathcal{Q}}}(v) = r(v) + f_{u_{\mathcal{Q}}}(v) - b_{u_{\mathcal{Q}}}(u_j, v)$$

on $\mathcal{Q}_j^\circ$, we can rewrite the linear variational problems (1.20) and (1.23) as

$$e_{\mathcal{Q}} \in \mathcal{Q}_j^\circ : \quad a_{u_{\mathcal{Q}}}(e_{\mathcal{Q}}, v) = r_{u_{\mathcal{Q}}}(v), \quad \forall v \in \mathcal{Q}_j^\circ, \tag{1.24}$$

and

$$\tilde{e}_{\mathcal{Q}} \in \mathcal{Q}_j^\circ : \quad \tilde{a}_{u_{\mathcal{Q}}}(\tilde{e}_{\mathcal{Q}}, v) = r_{u_{\mathcal{Q}}}(v), \quad \forall v \in \mathcal{Q}_j^\circ. \tag{1.25}$$

Observe that (1.22) can be rewritten as $r_{u_{\mathcal{Q}}}(v) = 0$, $\forall v \in \mathcal{S}_j^\circ$.

Using (P1)–(P3), we have traced back our original *nonlinear* defect problems to the corresponding *linear* case. Following Deuflhard, Leinen and Yserentant [45] p. 16, we now introduce the hierarchical splitting

$$\mathcal{Q}_j^\circ = \mathcal{S}_j^\circ \oplus \mathcal{V}_{\mathcal{Q}}$$

with $\mathcal{V}_{\mathcal{Q}} \subset \mathcal{V}_{\mathcal{E}} = \text{ span } \{\lambda_p^{\mathcal{Q}} \mid p \in \mathcal{N}_{\mathcal{E}}\}$. Here, we made use of condition (P4). The hierarchical splitting gives rise to a corresponding decomposition of both global problems (1.24) and (1.25) on $\mathcal{Q}_j^\circ$ in subproblems on $\mathcal{S}_j^\circ$ and $\mathcal{V}_{\mathcal{Q}}$, respectively. Now the assertion follows from a slight modification of the arguments in [45]. We refer to Kornhuber [84] for details. □

Of course, the conditions (P1)–(P4) are fulfilled if $\phi \equiv 0$, but can be hardly expected to hold for reasonable nonlinear problems. Nevertheless, we found very satisfying effectivity rates in our numerical calculations reported below (see also Kornhuber [84, 85]). A theoretical explanation of these observations will be the subject of future research.

4.1.3 An Error Estimate Based on Nonlinear Iteration

In order to increase the robustness of the error estimate, we will now incorporate additional *low–frequency* contributions in the localization step. For this reason, we consider a general *nonlinear iterative scheme*

$$e^{\nu+1} = e^\nu + \mathcal{B}(e^\nu), \quad \nu = 0, 1, \ldots, \tag{1.26}$$

which is intended to play the role of a (nonlinear) preconditioner. Of course, the evaluation of $\mathcal{B}$ should have optimal numerical complexity.

One step of the iteration (1.26) applied to the initial iterate 0 provides the approximation

$$\tilde{e}_{\mathcal{Q}} = \mathcal{B}(0). \tag{1.27}$$

The following proposition is an immediate consequence of the triangle inequality.

Proposition 4.4 *Assume that the iteration (1.26) is convergent with convergence rates bounded by* ρ, $0 \leq \rho < 1$. *Then we have the estimates*

$$(1+\rho)^{-1}\|\tilde{e}_{\mathcal{Q}}\| \leq \|e_{\mathcal{Q}}\| \leq (1-\rho)^{-1}\|\tilde{e}_{\mathcal{Q}}\|. \tag{1.28}$$

As a consequence of Propositions 4.1 and 4.4, a convergent iterative solver for the discrete defect equation (1.2) provides lower and upper bounds of the approximation error. The quality of the error estimate relies heavily on the convergence rates. This motivates the extension of monotone multigrid methods to the piecewise quadratic case.

Let us briefly consider extended underrelaxations for the discrete minimization problem (1.2) based on piecewise quadratic finite elements. Assume that each sequence of search directions $(M^\nu)_{\nu\geq 0} \subset \mathcal{Q}_j$ starts with the quadratic nodal basis $\Lambda_j^{\mathcal{Q}}$ (cf. condition (M1), p. 51). Then extended relaxations and underrelaxations induced by $(M^\nu)_{\nu\geq 0}$ are defined in analogy to (2.1.23) and (2.1.28), respectively. Convergence results can be derived by the same arguments as used in the proofs of Theorems 2.1 and 2.2.

Extending the quadratic nodal basis $\Lambda_j^{\mathcal{Q}}$ by the multilevel nodal basis $\Lambda_{\mathcal{S}}$ for all $\nu \geq 0$, we obtain the constant sequence $M^\nu = \Lambda_{\mathcal{Q}}$,

$$\Lambda_{\mathcal{Q}} = (\Lambda_j^{\mathcal{Q}}, \Lambda_{\mathcal{S}}),$$

of search directions. Replacing Λ by $\Lambda_{\mathcal{Q}}$, we can derive globally convergent standard and truncated monotone multigrid methods almost literally in the same way as in the previous chapter. In particular, the definition of quasioptimal monotone restrictions can be left unchanged. Note that this approach leads to nonlinear versions of well–known defect–correction schemes (cf. e.g. Aunzinger and Stetter [5] and Hackbusch [65]).

Assume that an implementation of monotone multigrid methods is available for piecewise linear finite elements. Then we only have to add a *fine grid smoother* on $\mathcal{Q}_j$ in order to get related solvers for the piecewise quadratic case. The evaluation of the stiffness matrix and the right–hand side can be partly performed by interpolation of the related piecewise linear data which have been already computed.

We apply one step of a *truncated monotone multigrid method with symmetric smoother* to (1.2) with the initial iterate $w_0 = 0$ in order to compute the intermediate iterates w_l, $l = 1, \ldots, \tilde{m}_{\mathcal{Q}}$, and the final approximate correction

$$\tilde{e}_{\mathcal{Q}} = \mathcal{B}(0) = w_{\tilde{m}_{\mathcal{Q}}}.$$

In order to apply our results on upper bounds for the *asymptotic convergence rates* (cf. Theorems 3.10, 3.11), we require the invariance of the discrete phases.

(I) The discrete phases of all intermediate iterates $\tilde{u}_j + w_l$, $l = 0, \dots, \tilde{m}_{\mathcal{Q}}$, coincide with the discrete phases of $u_{\mathcal{Q}}$.

Compare the similar condition (P2) stated in the previous section.

Theorem 4.5 *Assume that the condition (I) is satisfied. Then we have the estimates*

$$\tfrac{1}{2}\|\tilde{e}_{\mathcal{Q}}\| \le \|e_{\mathcal{Q}}\| \le C(j+2)^3\|\tilde{e}_{\mathcal{Q}}\|. \tag{1.29}$$

The constant C depends only on the ellipticity of $a(\cdot,\cdot)$, on the maximal coefficient b_i, $i = 0, \dots, N$, from (V3)' on p. 23, and on the initial triangulation $\mathcal{T}_0$.

Proof. As a consequence of condition (I), the nonlinear multigrid method reduces to a *linear* iteration for the variational equality (1.24) with u_j replaced by $\tilde{u}_j$. Hence, we can apply the general convergence theory for successive subspace corrections as condensed by Yserentant [126] to derive an upper bound for the asymptotic convergence rate of this scheme.

Symmetric point Gauß–Seidel smoothing is successively applied on the subspaces $\mathcal{W}_{\mathcal{Q}} = \mathcal{Q}_j^\circ$ and

$$\mathcal{W}_k = \text{ span } \{T_{\mathcal{Q},k}\lambda_p^{(k)} \mid p \in \mathcal{N}_k\}, \quad k = 0, \dots, j.$$

Here, the truncation operators $T_{\mathcal{Q},k} : \mathcal{Q}_j \to \mathcal{Q}_j^\circ$ are defined by

$$T_{\mathcal{Q},k} = I_{\mathcal{Q}_j^\circ} I_{\mathcal{S}_{\mathcal{Q},k}^\circ} \dots I_{\mathcal{S}_{\mathcal{Q},0}^\circ},$$

where $I_{\mathcal{Q}_j^\circ} : \mathcal{Q}_j \to \mathcal{Q}_j^\circ$ and $I_{\mathcal{S}_{\mathcal{Q},k}^\circ} : \mathcal{Q}_j \to \mathcal{S}_{\mathcal{Q},k}^\circ$ denote the interpolation to $\mathcal{Q}_j^\circ$ and $\mathcal{S}_{\mathcal{Q},k}^\circ$, respectively. The reduced spaces $\mathcal{S}_{\mathcal{Q},k}^\circ$ are defined by

$$\mathcal{S}_{\mathcal{Q},k}^\circ = \{v \in \mathcal{S}_k \mid v(p) = 0,\ \forall p \in \mathcal{N}_k \cap \mathcal{N}_{\mathcal{Q}}^\bullet\}$$

with $\mathcal{N}_{\mathcal{Q}}^\bullet$ introduced on p. 101. Observe the analogy to the definition (3.2.30) of the truncation operators $T_{j,k}^\nu$.

The symmetric Gauß–Seidel smoothing on $\mathcal{W}_\mathcal{Q}$ and $\mathcal{W}_k$, $k = 0, \ldots, j$, is represented by the scalar products $b_\mathcal{Q}(\cdot,\cdot)$ and $b_k(\cdot,\cdot)$, $k = 0, \ldots, j$, respectively. The modified interpolation operator $I_\mathcal{Q} : \mathcal{Q}_j \to \mathcal{S}_j$,

$$(I_\mathcal{Q} v)(p) = \begin{cases} v(p), & \text{if } \lambda_p^{(j)} \in \mathcal{Q}_j^\circ, \\ 0, & \text{otherwise,} \end{cases} \qquad p \in \mathcal{N}_j,$$

generates the direct splitting

$$\mathcal{Q}_j^\circ = \mathcal{V}_\mathcal{S} \oplus \mathcal{V}_\mathcal{Q} \tag{1.30}$$

into the subspaces $\mathcal{V}_\mathcal{S} = I_\mathcal{Q}\mathcal{Q}_j^\circ \subset \mathcal{S}_j$ and $\mathcal{V}_\mathcal{Q} = (id - I_\mathcal{Q})\mathcal{Q}_j^\circ \subset \mathcal{W}_\mathcal{Q}$. It can be shown by suitable scaling arguments (cf. Xu [122], Yserentant [126]) that the two–level splitting $v = v_\mathcal{S} + v_\mathcal{Q}$ with $v_\mathcal{S} \in \mathcal{V}_\mathcal{S}$ and $v_\mathcal{Q} \in \mathcal{V}_\mathcal{Q}$ is stable in the sense that

$$c\left(\|v_\mathcal{S}\|^2 + b_\mathcal{Q}(v_\mathcal{Q}, v_\mathcal{Q})\right) \le \|v\|^2 \le C\left(\|v_\mathcal{S}\|^2 + b_\mathcal{Q}(v_\mathcal{Q}, v_\mathcal{Q})\right) \tag{1.31}$$

holds for all $v \in \mathcal{Q}_j^\circ$. The hierarchical splitting

$$\mathcal{V}_\mathcal{S} = \bigoplus_{k=0}^{j} \mathcal{V}_k, \quad \mathcal{V}_k \subset \mathcal{W}_k,\ k = 0, \ldots, j,$$

of $\mathcal{V}_\mathcal{S} \subset \mathcal{S}_j$ is generated by modified interpolation operators as defined in (3.3.39).

Using Proposition 3.7, we get the estimate

$$\sum_{k=0}^{j} 4^k \|v_k\|_0^2 \le c(j+1)^2 \|\sum_{k=0}^{j} v_k\|^2, \quad \forall v_k \in \mathcal{V}_k,\ k = 0, \ldots, j. \tag{1.32}$$

Here, the L^2–norm $\|\cdot\|_0$ on the reduced domain is defined in analogy to Section 3.3.1, p. 84. It is well–known (cf. e.g. Yserentant [126], p. 298) that we have

$$c_1 b_k(v_k, v_k) \le 4^k \|v_k\|_0^2 \le C_1 b_k(v_k, v_k),\ \forall v_k \in \mathcal{V}_k,\ k = 0, \ldots, j, \tag{1.33}$$

for the bilinear forms induced by symmetric point Gauß–Seidel smoothing. The combination with (1.32) and (1.31) yields

$$\sum_{k=0}^{j} b_k(v_k, v_k) + b_{\mathcal{Q}}(v_{\mathcal{Q}}, v_{\mathcal{Q}}) \le c(j+2)^2 \| \sum_{k=0}^{j} v_k + v_{\mathcal{Q}} \|, \tag{1.34}$$

for all $v_k \in \mathcal{V}_k$, $k = 0, \ldots, j$, and all $v_{\mathcal{Q}} \in \mathcal{V}_{\mathcal{Q}}$.

On the other hand, a strengthened Cauchy–Schwarz inequality and an inverse inequality lead to

$$\| \sum_{k=0}^{j} v_k \|^2 \le C \sum_{k=0}^{j} 4^k \|v_k\|_0^2, \quad \forall v_k \in \mathcal{V}_k,\ k = 0, \ldots, j.$$

Similar arguments are used for example by Bornemann and Yserentant [29] or Yserentant [126]. Together with (1.33) and (1.31), this implies the upper estimate

$$\| \sum_{k=0}^{j} v_k + v_{\mathcal{Q}} \|^2 \le C \sum_{k=1}^{j} b_k(v_k, v_k) + b_{\mathcal{Q}}(v_{\mathcal{Q}}, v_{\mathcal{Q}}),$$

for all $v_k \in \mathcal{V}_k$, $k = 0, \ldots, j$, and all $v_{\mathcal{Q}} \in \mathcal{V}_{\mathcal{Q}}$.

Now the upper bound

$$\rho = 1 - c(j+2)^{-3} < 1$$

for the asymptotic convergence rate follows from Theorem 5.4 in the overview of Yserentant [126], and the assertion then is an immediate consequence of Proposition 4.4. □

Again, the exponent of j can be improved in more regular situations. The condition (I) is less restrictive than (P1)–(P4), indicating that error estimates based on monotone multigrid methods are more robust than the error estimate based on diagonal scaling. However, in comparison with the evaluation of (1.18), one step of a multigrid method is about twice as expensive.

4.2 A Stopping Criterion for the Adaptive Algorithm

Assume that we have computed an a posteriori error estimate $\| \tilde{e}_{\mathcal{Q}} \|$, where $\| \cdot \|$ either stands for the diagonal scaling $| \cdot |$ (cf. Theorem 4.3) or for the energy norm $\| \cdot \|$ (cf. Theorem 4.5). We will now use this estimate to provide a stopping criterion for the complete adaptive algorithm. In order to compensate a possible underestimation of the true approximation error, we introduce a safety factor σ_{app} with $0 < \sigma_{app} \leq 1$. The adaptive algorithm is stopped if the condition

$$\| \tilde{e}_{\mathcal{Q}} \| \leq \sigma_{app} \mathrm{TOL} \tag{2.35}$$

is satisfied. If a *relative tolerance*, say $\mathrm{TOL} = \varepsilon \|u\|$, is prescribed, then the unknown norm $\|u\|$ is replaced by the actual approximation $\|\tilde{u}_j\|$.

The efficiency and reliability of the whole adaptive algorithm may depend heavily on the choice of the safety factor σ_{app}. While a lot of computational work is wasted, if σ_{app} is chosen too small, the desired accuracy may be missed in the opposite case. Up till now, there is no sound mathematical insight how this parameter should be determined. We will select $\sigma_{app} = 1.$ in our numerical experiments which will turn out to be a reasonable choice for well–behaved problems but might be dangerous in other situations. As a consequence, the quality of the numerical solution of real–life problems is usually checked by additional heuristic criteria based on the special physical situation. This is beyond the scope of our general discussion.

4.3 Error Indicators and Local Refinement

We want to refine the triangulation in such a way that the *discretization error* $\|u - u_j\|$ is efficiently reduced. This strategy simultaneously provides an approximation of the phases of the solution u, provided that the accuracy deteriorates in the neighborhood of the free boundary. Due to a loss of regularity, this is frequently the case. However, the adaptive strategy to be presented should be modified if a highly accurate approximation of the different phases is the main issue of the computation. In this case, we may additionally refine all triangles neighboring the free boundary in each refinement step.

We assume that the algebraic error is small enough. Based on the *global* estimate $\| \tilde{e}_{\mathcal{Q}} \|$ of the approximation error, *local* error indicators to control the adaptive refinement process are selected as follows.

Consider the hierarchical splitting $\mathcal{Q}_j = \mathcal{S}_j \oplus \mathcal{V}_{\mathcal{E}}$ with

$$\mathcal{V}_{\mathcal{E}} = \text{ span } \{\lambda_p^{\mathcal{Q}} \mid p \in \mathcal{N}_{\mathcal{E}}\}.$$

Note that $\mathcal{V}_{\mathcal{E}}$ consists of the quadratic bubble functions associated with the midpoints of the interior edges $\mathcal{E}_j$. We decompose $\tilde{e}_{\mathcal{Q}}$ according to

$$\tilde{e}_{\mathcal{Q}} = \tilde{e}_{\mathcal{S}} + \tilde{e}_{\mathcal{E}}$$

with $\tilde{e}_{\mathcal{S}} \in \mathcal{S}_j$ and $\tilde{e}_{\mathcal{E}} \in \mathcal{V}_{\mathcal{E}}$. Here, $\tilde{e}_{\mathcal{S}}$ and $\tilde{e}_{\mathcal{E}}$ represent the low– and high–frequency parts of $\tilde{e}_{\mathcal{Q}}$, respectively. It is easily checked that we have

$$c \left(\| \tilde{e}_{\mathcal{S}} \|^2 + |\tilde{e}_{\mathcal{E}}|^2 \right) \leq \| \tilde{e}_{\mathcal{Q}} \|^2 \leq C \left(\| \tilde{e}_{\mathcal{S}} \|^2 + |\tilde{e}_{\mathcal{E}}|^2 \right), \tag{3.36}$$

where

$$|\tilde{e}_{\mathcal{E}}|^2 = \sum_{p \in \mathcal{N}_{\mathcal{E}}} \eta_p, \qquad \eta_p = \tilde{e}_{\mathcal{E}}(p)^2 a(\lambda_p^{\mathcal{Q}}, \lambda_p^{\mathcal{Q}}), \;\; p \in \mathcal{N}_{\mathcal{E}}. \tag{3.37}$$

We want to refine the given triangulation in such regions where the high–frequency contributions deteriorate the overall accuracy. Assuming that the high frequency part of the *discretization error* $\|u - u_j\|^2$ is represented by $|\tilde{e}_{\mathcal{E}}|^2$, the *local contributions* η_p, $p \in \mathcal{N}_{\mathcal{E}}$, are used as *local error indicators*.

A triangle $t \in \mathcal{T}_j$ is marked for refinement, if at least one of the indicators η_p associated with the edges of t exceeds a certain threshold $\sigma_{ref}\bar{\eta}$. Here $\bar{\eta}$ is a guess of the maximal local error arising on the next level in case of uniform refinement and $\sigma_{ref} < 1$ is a safety factor. In the numerical examples reported below, $\bar{\eta}$ is computed by local extrapolation (cf. Babuška and Rheinboldt [6]) and we chose $\sigma_{ref} = 0.5$. In the case of linear selfadjoint problems, a theoretical justification of a similar approach was recently given by Dörfler [46].

Now assume that a subset $\bar{\mathcal{T}}_j \subset \mathcal{T}_j$ of triangles has been marked for refinement. To preserve the shape regularity of the initial triangulation, irregular triangles must not be refined further. Hence, all green refinements are skipped replacing the irregular triangles contained in $\bar{\mathcal{T}}_j$ by their fathers. After the regular refinement of all $t \in \bar{\mathcal{T}}_j$ there may exist triangles with edges

which are refined twice or with two or more bisected edges. Regular refinement is continued until no such triangles are left. The remaining irregular vertices are remedied by green closure.

In general, this dynamic refinement process will not produce a nested sequence of triangulations in the sense of the conditions (T1)–(T3) stated in Section 3.1.1. However, a nested sequence $\mathcal{T}_0, \ldots, \mathcal{T}_j$ can be uniquely reconstructed from the initial triangulation $\mathcal{T}_0$ and the actual triangulation $\mathcal{T}_j$ (with refinement depth j) alone. We emphasize that this is possible without additional computational effort, if the underlying data structures are properly chosen. We refer to Bank [8], Leinen [90], Beck, Erdmann and Roitzsch [19], Bastian [17], and others. Note that the (non–negative) difference of refinement level and refinement depth can be used to judge the quality of the underlying refinement strategy.

4.4 A Stopping Criterion for the Iterative Solver

The iterative solution of the discrete problem (1.3.57) should be stopped as soon as the *algebraic error* $\|u_j - u_j^\nu\|$ is small enough. Recall for example the accuracy condition (1.8).

We start with the observation that the corrections resulting from a good iterative method provide good a posteriori estimates for the algebraic error. Similarly, preconditioned residuals are frequently used in the linear self-adjoint case (cf. eg. Deuflhard, Leinen and Yserentant [45], Bornemann, Erdmann and Kornhuber [27] or Becker, Johnson and Rannacher [20]).

Theorem 4.6 *Assume that the discrete problem (1.3.57) is non–degenerate in the sense of (2.3.42) and that the iterates $(u_j^\nu)_{\nu\geq 0}$, are computed by a monotone multigrid method. Then the a posteriori estimate*

$$\tfrac{1}{2}\|u_j^{\nu+1} - u_j^\nu\| \leq \|u_j - u_j^\nu\| \leq C(j+1)^s \|u_j^{\nu+1} - u_j^\nu\|, \ \forall \nu \geq \nu_0, \quad (4.38)$$

holds for sufficiently large $\nu_0 \geq 0$. We have $s = 4$ for the standard monotone multigrid method induced by Λ and $s = 6$ for the corresponding truncated version. The constant C depends only on the ellipticity of $a(\cdot,\cdot)$, on the maximal coefficient b_i, $i = 0, \ldots, N$, from (V3)' on p. 23, and on the initial triangulation $\mathcal{T}_0$.

Proof. Using Theorem 3.10 or Theorem 3.11, we can find a $\nu_0 \geq 0$ and a suitable positive constant c such that the error estimate

$$\|u_j - u_j^{\nu+1}\| \leq (1 - c(j+1)^{-s})\|u_j - u_j^{\nu}\|, \quad \forall \nu \geq \nu_0, \tag{4.39}$$

holds with $s = 4$ or $s = 6$, respectively. Now the assertion follows by the triangle inequality in the same way as Proposition 4.4. □

In the light of Theorem 4.6, the corrections $\|u_j^{\nu+1} - u_j^{\nu}\|$ provided by monotone multigrid methods will be used as a posteriori estimates for the algebraic error. Good initial iterates, as required in the proof of the asymptotic estimate (4.38), are usually obtained by interpolating the final iterate from the preceding level. Other values of s in (4.38) can be obtained by other variants of the asymptotic error estimate (4.39) (cf. Section 3.3).

Based on the above considerations, we now give a stopping criterion for the iterative solver. Recall that we want to solve the given continuous problem up to a prescribed tolerance TOL. With some safety factor $\sigma < 1$, the requirement

$$\|u_j - \tilde{u}_j\| \leq \sigma \mathrm{TOL}, \quad \forall j = 0, \ldots, J, \tag{4.40}$$

guarantees that the overall accuracy on the final triangulation $\mathcal{T}_J$ is not deteriorated by the algebraic error. On the preceding coarser levels $j = 0, \ldots, J-1$ the criterion (4.40) is intended to provide a *sufficient damping of the low frequencies* of the initial iterates $u_{j+1}^0 = \tilde{u}_j$.

Such *cascadic iterations* have been introduced by Deuflhard [44], who discovered that stopping criteria of the form (4.40) lead to a considerable stabilization of the iterative solver. For the linear selfadjoint case, a theoretical analysis was given by Shaidurov [112] and Bornemann and Deuflhard [26]. In particular, it was shown in [26] that, under suitable assumptions, the stopping criterion (4.40) gives usual single grid smoothers a multigrid speed of convergence.

Recall that our adaptive refinement strategy relies on the fact that the high–frequency contributions of the approximate correction $\tilde{e}_{\mathcal{Q}}$ reflect the behavior of the discretization error and are not caused by insufficient algebraic approximation. This is an additional motivation for the restrictive accuracy assumptions (4.40).

Ignoring constants, let us assume for the moment that our estimates are representing the algebraic and the approximation error exactly. Then the above

stopping criterion for the algebraic solver implies that the algebraic accuracy condition (1.8) holds with $\beta_a = \sigma/(1-\sigma)$, provided that the approximation error is greater than TOL, i.e. that the final level is not yet reached. On the final level, (1.8) still holds with $\beta_a = \sigma/(\frac{1}{2}-\sigma)$, if the approximation error has not been reduced by more than a factor of 2 in the final refinement step. This is a reasonable assumption, because asymptotically the discretization error is well–known to decrease at most linearly with the maximal stepsize which in turn can be only halved in each refinement step.

In practical computations, we have to approximate the algebraic error appearing in (4.40) by an a posteriori estimate: The iterate $\tilde{u}_j = u_j^{\nu_0+1}$ is accepted as soon as the *stopping criterion*

$$\|\tilde{u}_j - u_j^{\nu_0}\| \leq \sigma_{alg}\sigma_{app}\mathrm{TOL}. \tag{4.41}$$

is fulfilled. The new safety factor $\sigma_{alg} < 1$ is intended to compensate the approximation of the true algebraic error. We chose $\sigma_{alg} = 0.1$ for our numerical computations.

5 Numerical Results

In the preceding two chapters, we have constructed the building blocks of an adaptive multilevel method for our non–smooth minimization problem (1.2.33): iterative solvers, stopping criteria based on a posteriori error estimates and an adaptive refinement strategy. Apart from the accuracy TOL, the whole algorithm contains only three parameters: the safety factors σ_{app} and σ_{alg} for the approximation error and for the algebraic solution, respectively, and a safety factor σ_{ref} involved in the refinement strategy.

In the following numerical examples, we will always prescribe a relative tolerance of 5%. Hence, the algorithm stops as soon as $\| \tilde{e}_{\mathcal{Q}} \| \leq 0.05\sigma_{app}\|\tilde{u}_j\|$ is satisfied (cf. Section 4.2). Recall that $\| \tilde{e}_{\mathcal{Q}} \|$ is an a posteriori error estimate for the actual approximation $\tilde{u}_j$ of the solution u (see Section 4.1). On each refinement level j, the iterative solution of the discrete problems is continued until $\|u_j^{\nu+1} - u_j^{\nu}\| \leq 0.05\sigma_{alg}\sigma_{app}\|u_j^{\nu}\|$ is fulfilled (cf. Section 4.4). We will use the same default values $\sigma_{app} = 1.$, $\sigma_{alg} = 0.1$, and $\sigma_{ref} = 0.5$ in all the numerical examples to be reported in this chapter.

The implementation was carried out at the Konrad–Zuse–Center in Berlin in the framework of a recent C++ version of the finite element toolbox KASKADE developed by Beck, Erdmann and Roitzsch [19].

For further numerical results including comparisons with previous multigrid approaches, we refer to Kornhuber [82, 83, 85].

5.1 Deformation of a Membrane with a Rigid Obstacle

In our first example, we compute the vertical displacement u of a planar membrane Ω which is exposed to a force density f and constrained by a rigid upper obstacle φ. We know from Section 1.1.1 that u is the solution of the convex minimization problem (1.1.8).

Multigrid methods for such obstacle problems have been derived by Brandt and Cryer [33], Hackbusch and Mittelmann [66], Hoppe [69, 70, 71], Mandel [92, 93], and others. Roughly speaking, truncated monotone multigrid methods combine the global convergence of Mandel's method with the fast convergence speed of Brandt and Cryer's scheme. We refer to Kornhuber [82] for details. A complete adaptive multilevel method for obstacle problems was recently proposed by Hoppe and Kornhuber [75]. In contrast to the algorithm presented here, their iterative solver is based on active set strategies and the a posteriori error estimator is slightly different.

Recall that the obstacle problem (1.1.8) is a special case of our reference problem (1.2.33), or (for variable obstacle) can be transformed to such a problem (with constant obstacle) by a simple translation. We will apply our algorithm directly to (1.1.8), using the data

$$\varphi(x) = (4r)^3 - 1, \quad r = ((0.5 - x_1)^2 + (0.5 - x_2)^2)^{1/2},$$

$\alpha \equiv 1$, and $f \equiv 0$ on the unit square $\Omega = (0,1) \times (0,1)$.

5.1.1 The Adaptive Multilevel Method

We now employ our adaptive multilevel algorithm, using the *truncated* monotone multigrid method (cf. Theorem 3.11) and the *local* a posteriori error estimator (cf. Theorem 4.3). The initial triangulation is shown in Figure 5.1.

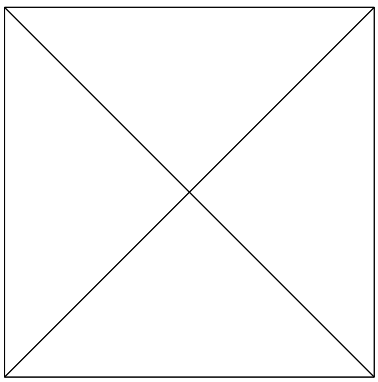

Figure 5.1 Initial triangulation $\mathcal{T}_0$

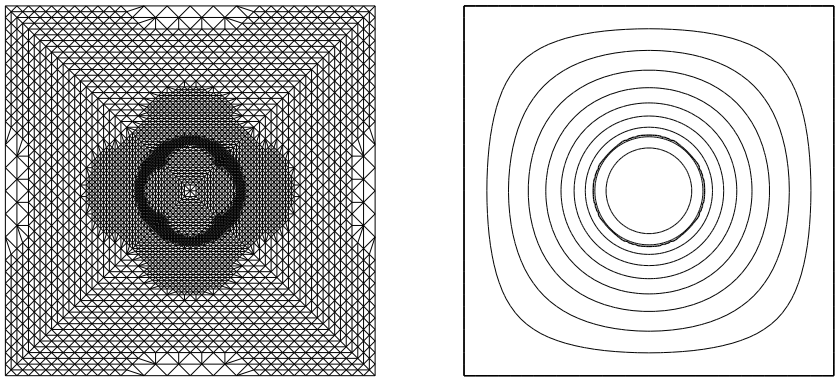

Figure 5.2 Final triangulation $\mathcal{T}_8$ and final approximation $\tilde{u}_8$

Starting with $\mathcal{T}_0$, the algorithm generates a sequence of adaptively refined triangulations $\mathcal{T}_1, \ldots, \mathcal{T}_8$ and of corresponding approximations $\tilde{u}_0, \ldots, \tilde{u}_8$. Disregarding our preceding notation, the subscripts from now on denote the number of refinement steps. The final triangulation $\mathcal{T}_8$ is depicted in the left picture of Figure 5.2. The right picture shows the level curves of the corresponding final approximation $\tilde{u}_8$. The (free) boundary of the coincidence set is represented by a bold line. Observe that the refinement mainly concentrates on the resolution of the contact region with special emphasis on the free boundary.

Level	Depth	Nodes	Iterations	Error %
0	0	1	2	57.7
1	1	5	2	37.7
2	2	25	3	78.7
3	3	41	3	37.1
4	4	169	4	18.0
5	5	533	3	10.1
6	6	1273	3	6.6
7	6	1665	2	5.6
8	7	3869	2	3.9

Table 5.1 Approximation history

The complete approximation history is reported in Table 5.1. On the first levels, the triangulation might be too coarse to guarantee the saturation assumption (4.1.7). Hence, the resulting error estimates should be handled with care. Nevertheless, the corresponding refinement indicators seem to work well. The a posteriori error estimates on the subsequent levels suggest that the error behaves like $\mathcal{O}(n_j^{-1/2})$. This is in agreement with well–known $\mathcal{O}(h)$ error estimates (see e.g. Ciarlet [39]) and illustrates that the sequence of triangulations can be regarded as quasioptimal in the sense that it reproduces the optimal order of approximation.

The number of iterations stays moderate throughout the approximation. Note that we always need one additional iteration to control the algebraic error. More information on the behavior of the monotone multigrid methods will be given in the next section.

5.1.2 The Monotone Multigrid Methods

Let us apply our multigrid methods to the discrete obstacle problem on $\mathcal{T}_8$, using the hierarchy of triangulations $\mathcal{T}_0, \ldots, \mathcal{T}_8$. The standard monotone multigrid method (cf. Theorem 3.10) and the truncated version (cf. Theorem 3.11) will be denoted by STDKH and TRCKH, respectively.

In our first experiment, we investigate the convergence behavior on the fixed refinement level $j = 8$, starting with the initial iterate $u_j^0 = 0$. Figure 5.3 shows the decrease of the algebraic errors in course of the iteration. The overall convergence behavior can be divided into a *transient* phase, dominated by the search for the (discrete) free boundary, and an *asymptotic* phase, corresponding to the iterative solution of the reduced linear problem. This observation supports the analysis contained in Section 2.3.

As compared to STDKH, the truncated version TRCKH exhibits a tremendous improvement of the asymptotic convergence rates giving a numerical justification for using truncated nodal basis functions. It is interesting that the transient convergence properties remain almost the same.

To provide a more realistic situation, the artificial initial iterate zero is now replaced by the interpolated solution from the previous level. In this way, the transient phase is completely eliminated from the convergence history as can be seen in Figure 5.4. In the beginning, STDKH and TRCKH are comparable in eliminating the high–frequency contributions of the error, but only TRCKH keeps this convergence speed throughout the whole iteration.

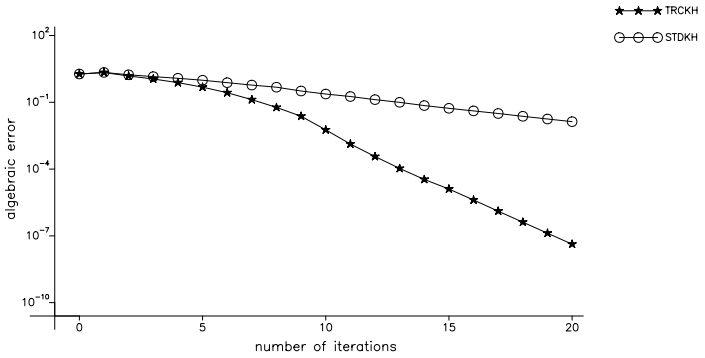

Figure 5.3 Iteration history: Initial iterate $u_j^0 = 0$

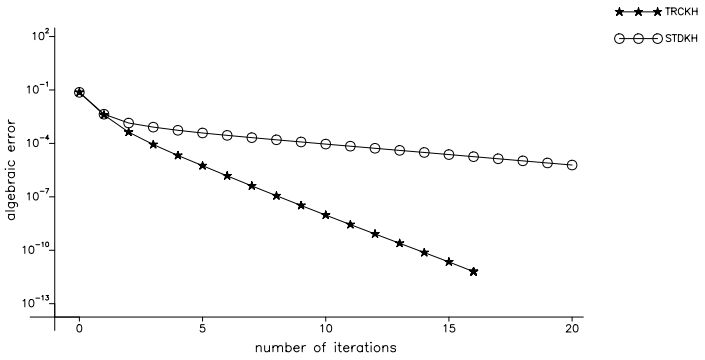

Figure 5.4 Iteration history: Interpolated initial iterate

To study the convergence properties for increasing j, we introduce the asymptotic efficiency rates ρ_j,

$$\rho_j = \sqrt[\nu_0]{\delta_j^{\nu_0}/\delta_j^0}, \quad j = 0, \ldots, 21, \tag{1.1}$$

where δ_j^ν denotes the algebraic error after ν iteration steps and the triangulations $\mathcal{T}_9, \ldots, \mathcal{T}_{21}$ are obtained by further adaptive refinement. We choose ν_0 such that $\delta_j^{\nu_0} < 10.^{-12}$. The results are shown in Figure 5.5. Observe the fast increase of the asymptotic efficiency rates of STDKH on level 4, reflecting the poor representation of the low–frequency contributions of the error.

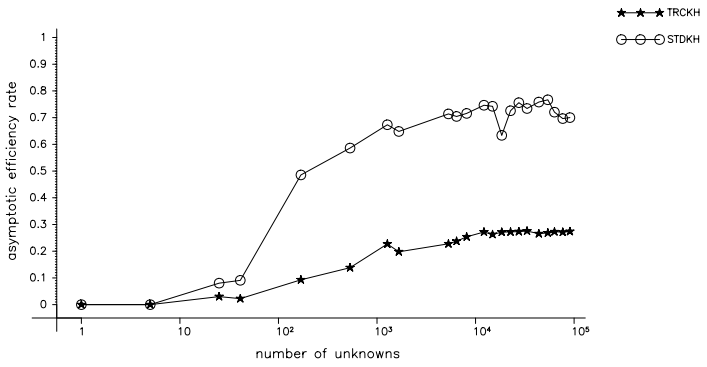

Figure 5.5 Asymptotic efficiency rates

However, the asymptotic efficiency rates of both multigrid methods seem to saturate with increasing j. It is not astonishing that our asymptotic estimates contained in the Theorems 3.10 and 3.11 are too pessimistic as compared with these calculations, because the regularity of the actual problem did not enter our theoretical considerations. Starting with $u_j^0 = 0$, the number of transient iteration steps becomes larger with increasing j. However, the transient convergence rates also seem to be uniformly bounded, suggesting the existence of *global* bounds for the convergence rates. A theoretical verification of these experimental results will be the subject of future research.

5.1.3 The A Posteriori Error Estimates

We now give a comparison of the local error estimate (cf. Theorem 4.3), as used above, and of an error estimate based on nonlinear iteration (cf. Theorem 4.5). As a nonlinear iteration, we chose the truncated monotone multigrid method with symmetric smoother (cf. Section 4.1.3).

The quality of the a posteriori error estimates $\| \tilde{e}_{\mathcal{Q}} \|$ of the approximation error $\|u - \tilde{u}_j\|$ on refinement level j is measured by the *effectivity indices*

$$\kappa_j = \| \tilde{e}_{\mathcal{Q}} \| / \|u - \tilde{u}_j\|, \quad j = 1, \ldots, 8. \tag{1.2}$$

We will replace the exact solution u appearing in (1.2) by the approximation $\tilde{u}$ resulting from two uniform refinements of the final triangulation $\mathcal{T}_8$.

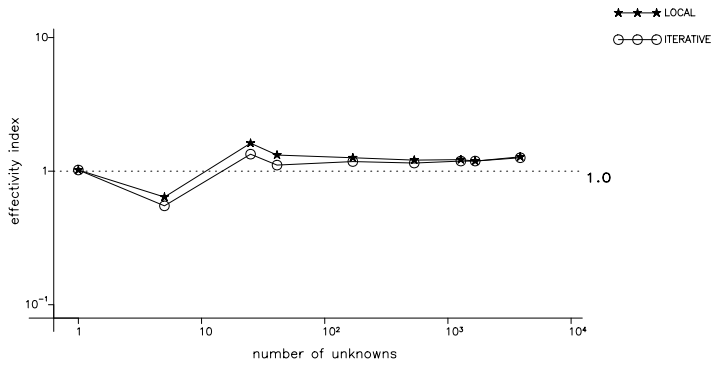

Figure 5.6 Effectivity indices of the approximation error estimates

Figure 5.6 shows the effectivity indices for the local and the iterative error estimate, denoted by LOCAL and ITERATIVE, respectively. Both a posteriori error estimators work satisfactory throughout the approximation. Hence, in this example, the additional computational effort for the iterative error estimate does not pay off. Note that the local error indicators derived from the iterative estimate also provide almost the same sequence of triangulations as obtained above.

We finally check the stopping criterion for the iterative solution based on the estimate $\|\tilde{u}_j - u_j^{\nu_0}\|$, $\tilde{u}_j = u_j^{\nu_0+1}$, of the algebraic error $\|u_j - u_j^{\nu_0}\|$. The performance of this error estimate relies heavily on the quality of the applied iterative solver (cf. Theorem 4.6). In the light of the fast convergence of the truncated monotone multigrid method, the resulting *algebraic effectivity indices*

$$\kappa_j = \|\tilde{u}_j - u_j^{\nu_0}\| / \|u_j - u_j^{\nu_0}\|, \qquad j = 0, \ldots, 8, \tag{1.3}$$

can be expected to be close to 1. In fact, we found κ_j ranging from 0.91 to 1.0. Using the standard version, we sometimes observed a slight underestimation of the true error with κ_j taking values from 0.39 to 1.0. Note that an underestimation of $\|u_j - u_j^{\nu_0}\|$ can still be compensated by the final correction. In this example, we found that, for both monotone multigrid methods, the algebraic error of the resulting final iterate $\tilde{u}_j = u_j^{\nu_0+1}$ was always less than the desired tolerance $0.05\sigma_{alg}\sigma_{app}\|u_j\|$.

5.2 A Strongly Reverse Biased p-n Junction

Though the membrane problem is very popular as an introductory example, it scarcely occurs in real–life applications. In order to illustrate the importance of obstacle problems from a more practical point of view, we now turn to a problem from semiconductor device simulation.

We consider a device occupying the domain $\Omega \subset \mathbb{R}^2$ whose stationary behavior can be described by the well–known drift–diffusion equations (see e.g. van Roosbroeck [107])

$$\begin{aligned} -\nabla \cdot (\varepsilon \nabla u) &= q(N - n + p), \\ \nabla \cdot J_n &= qR, & J_n &= q(D_n \nabla n - \mu_n n \nabla u), \\ \nabla \cdot J_p &= -qR, & J_p &= -q(D_p \nabla p - \mu_p p \nabla u), \end{aligned} \tag{2.4}$$

where usually the electric potential u and the carrier concentrations n and p for electrons and holes are unknown, while the permittivity ε, the doping profile N, the elementary charge q, the electron and hole diffusivities D_n and D_p, the electron and hole mobilities μ_n and μ_p, and the generation–recombination rate R are given parameters of the problem. The boundary $\partial\Omega$ consists of (ohmic) contacts $\partial\Omega_D = \partial\Omega_a \cup \partial\Omega_c$ and insulating segments $\partial\Omega_N$. This is reflected by Dirichlet boundary conditions for u, n, and p on $\partial\Omega_D$ and vanishing electric field $-\nabla u$ and current densities J_n, J_p on $\partial\Omega_N$. There is a vast literature on the merits and limits of this model. We refer for example to Selberherr [111] or Markowich, Ringhofer and Schmeiser [94]. It will turn out that the nonlinear system (2.4) can be considerably simplified under strongly reverse bias conditions.

Let us consider a p-n junction separating a p–region (where $N < 0$) from a n–region (where $N > 0$) as shown for example in Figure 5.7. For the moment, we ignore the insulating oxide region on the top of the device. Assume that the reverse voltage $-u_a$ is applied at the anode, while the voltage at the cathode is kept zero. Then the carriers are driven away from the neighborhood of the p-n junction leaving a *depletion area* Ω_d, where ideally no carriers are present. In Ω_d the potential u is bounded by the applied voltage so that we have

$$n = p = 0, \quad -u_a < u < 0 \qquad \text{in } \Omega_d. \tag{2.5}$$

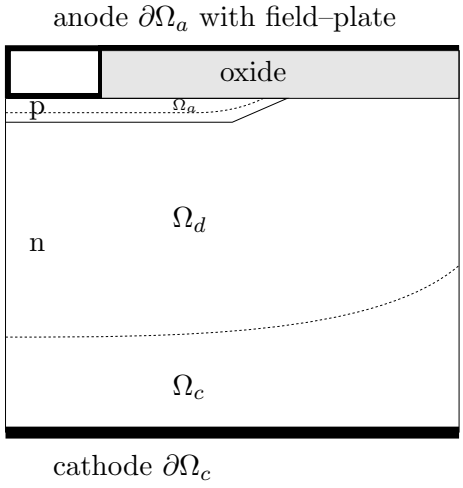

Figure 5.7 A reverse biased p-n junction

The depletion area separates the remaining parts Ω_a and Ω_c of the silicon where u is equal to $-u_a$ and 0, respectively. For large u_a the *total depletion assumption*

$$\begin{aligned} &n = 0, \quad p = -N, \quad \nabla u = 0 \qquad \text{in } \Omega_a, \\ &n = N, \quad p = 0, \quad \nabla u = 0 \qquad \text{in } \Omega_c, \end{aligned} \tag{2.6}$$

holds. Inserting (2.5) and (2.6) in the drift–diffusion equations (2.4), we obtain

$$\begin{aligned} u &= -u_a && \text{in } \Omega_a, \\ -\nabla \cdot (\varepsilon \nabla u) &= qN && \text{in } \Omega_d, \\ u &= 0 && \text{in } \Omega_c. \end{aligned} \tag{2.7}$$

It is reasonable to assume that the potential u and the electric displacement $-\varepsilon\nabla u$ are continuous across the interior (free) boundaries $\Gamma_a = \Omega_a \cap \bar{\Omega}_d$ and $\Gamma_c = \Omega_c \cap \bar{\Omega}_d$, giving

$$\begin{aligned} &u = -u_a, \quad \nabla u \cdot n_{\Gamma_a} = 0 \qquad \text{on } \Gamma_a, \\ &u = 0, \quad \nabla u \cdot n_{\Gamma_c} = 0 \qquad \text{on } \Gamma_c, \end{aligned} \tag{2.8}$$

with n_{Γ_a} and n_{Γ_c} denoting normals on Ω_a and Ω_c, respectively. Because there are no carriers in the oxide anyway, we only have to require the continuity of u and $-\varepsilon\nabla u$ across the interface in order to extend this model to the whole device. Finally, u must satisfy the boundary conditions mentioned above.

We have derived the classical formulation of a double obstacle problem for the linear elliptic operator $-\nabla\cdot(\varepsilon\nabla\cdot)$ with the constant obstacles $-u_a$ and 0 (see e.g. Rodrigues [105]). The corresponding weak formulation is given by our reference problem (1.2.33) setting

$$a(v,w)=\int_\Omega \varepsilon\nabla v\cdot\nabla w\,dx,\quad \ell(v)=q\int_\Omega Nv\,dx,$$

and $\Phi=\chi_{[-u_a,0]}$ is the characteristic function of the interval $[-u_a,0]$. The solution space is given by $H=\{v\in H^1(\Omega)\mid v|_{\partial\Omega_a}=u_a,\ v|_{\partial\Omega_c}=0\}$.

This simplified model was proposed by Hunt and Nassif [76]. Using an appropriate scaling, the drift–diffusion equations (2.4) become singularly perturbed as $u_a\to\infty$. Then the simplified model is recovered as the corresponding reduced problem. We refer to Markowich, Ringhofer and Schmeiser [94] for further information.

Large peaks of the electric field may cause impact ionization which in turn leads to an avalanche breakdown in the device. In order to improve the blocking capability, high voltage p-n junctions are often equipped with multistep field–plates (see e.g. Feiler and Gerlach [54]). The most time–consuming part in the optimal geometrical design of such field–plates is the numerical solution of the obstacle problem. This motivates the application of fast solvers. As the electric field $-\nabla u$ and not u itself is of primary interest in this application, mixed methods might be an interesting subject of future research (see e.g. Wohlmut [121] for the linear selfadjoint case).

In the following numerical example, we will concentrate on the geometry depicted in Figure 5.7. The height of the device is $160\mu m$ and we apply a reverse voltage of $-u_a=-800V$ at $\partial\Omega_a$. The doping concentration has the values $N=-10^{17}cm^{-3}$ in the p–region and $N=8\cdot10^{13}cm^{-3}$ in the n–region, respectively. $q=1.602\cdot10^{-19}As$ is the elementary charge. The permittivity is given by $\varepsilon=\varepsilon_0\varepsilon_r$ with $\varepsilon_0=8.854\cdot10^{-14}As/Vcm$ and we have $\varepsilon_r=11.7$ in the silicon and $\varepsilon_r=3.9$ in the oxide, respectively. Note that in most real–life applications the p–region is about ten times thinner and the steps of the field plate are much lower. We will come back to this point later on.

5.2.1 The Adaptive Multilevel Method

We employ the same adaptive algorithm as in the previous example, using the truncated monotone multigrid method and the local a posteriori error estimate. The initial triangulation is depicted in Figure 5.8.

After 9 adaptive refinement steps, the algorithm has produced the final triangulation $\mathcal{T}_9$ which is shown in the left picture of Figure 5.9. The right picture illustrates the final solution $\tilde{u}_9$. Again the free boundaries are marked by

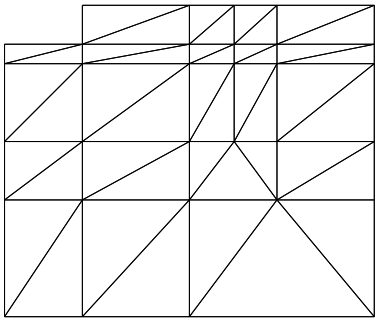

Figure 5.8 Initial triangulation $\mathcal{T}_0$

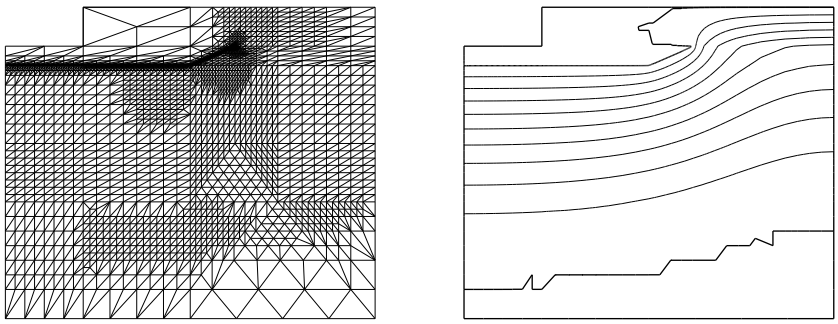

Figure 5.9 Final triangulation $\mathcal{T}_9$ and final approximation $\tilde{u}_9$

bold lines. Observe that the level curves of $\tilde{u}_9$ reflect the jump of the electric field resulting from the jumping permittivity across the silicon/oxide interface. The refinement concentrates on the peak of the electric field at the end

of the p-region while the free boundaries are only roughly approximated. This is in accordance with the construction of the error indicators which are intended to provide an efficient reduction of the energy error. In this example, such behavior is desired from the physical point of view. While the peaks of the electric field give some information on the breakdown voltage and thus have to be resolved properly, the actual free boundary is of minor interest. This may be different in other examples and we will come back to this point later on.

Level	Depth	Nodes	Iterations	Error %
0	0	21	8	32.3
1	1	74	4	21.3
2	2	178	4	16.0
3	3	311	3	13.3
4	4	424	3	10.7
5	5	870	3	7.9
6	6	1193	2	6.7
7	7	1538	2	5.8
8	7	2193	2	5.1
9	7	2796	2	3.2

Table 5.2 Approximation history

The complete approximation history is reported in Table 5.2. On the first refinement levels, we need some more iterations because the nonlinear Gauß–Seidel method is no longer an exact solver on the initial grid. Later on (where each step is 100 times more expensive), the number of iterations is as moderate as in our first example. Again the estimated approximation error seems to behave like $\mathcal{O}(n_j^{-1/2})$. Only in the beginning and on the final level do we get a slightly faster reduction than expected.

5.2.2 The Monotone Multigrid Methods

As in our previous example, we first apply the truncated monotone multigrid method (TRCKH) and the standard version (STDKH) to the discrete problem arising on the final level $j = 9$.

The iteration history corresponding to the initial iterate $u_j^0 = 0$ is shown in Figure 5.10. For both methods, the asymptotic convergence speed is almost

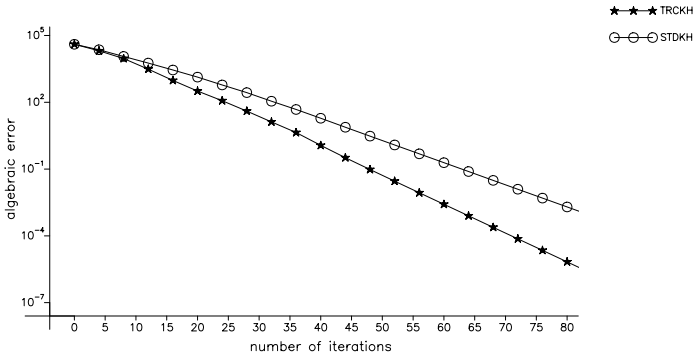

Figure 5.10 Iteration history: Initial iterate $u_j^0 = 0$

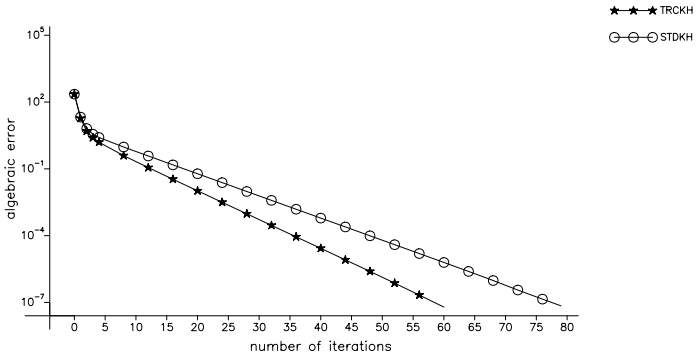

Figure 5.11 Iteration history: Interpolated initial iterate

4 times slower than above (cf. Figure 5.3). Starting with the interpolated initial iterate from the previous refinement level, we obtain the iteration history depicted in Figure 5.11. After a fast reduction of the high–frequency contributions, the iterations enter the asymptotic phase directly.

In order to illustrate the convergence behavior for increasing refinement, we consider the asymptotic efficiency rates ρ_j, $j = 0, \ldots, 24$, defined according to (1.1). The results are shown in Figure 5.12. This time, the effect of truncating the search directions is not as drastic as in our first example (cf. Figure 5.5). Again the asymptotic efficiency rates seem to saturate with increasing j.

All together, the constraints seem to be of minor importance for the whole

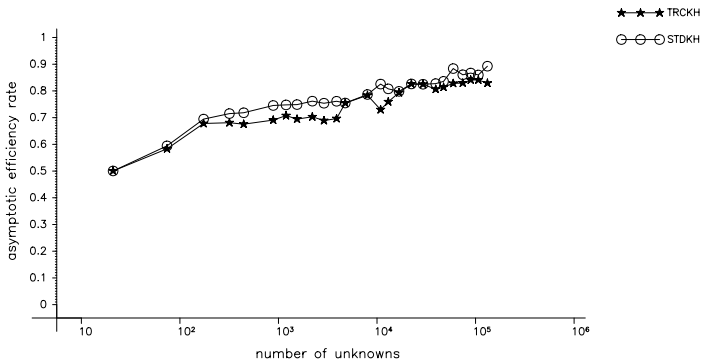

Figure 5.12 Asymptotic efficiency rates

iterative solution process. In fact, neglecting the obstacles and using a standard linear multigrid method with an exact solver on the coarse grid, we observed similar unsatisfying convergence rates. The reason is the geometry of Ω. To resolve the flat p–region, we either have to accept a large number of triangles in $\mathcal{T}_0$ or bad aspect ratios. Both properties are well–known to deteriorate the convergence speed of multigrid methods even in the linear selfadjoint case. We cannot expect monotone multigrid methods to do better. It should be mentioned that the p–region is usually much thinner and the resulting geometric difficulties are more severe. A possible remedy is *blue refinement* (cf. Kornhuber and Roitzsch [86, 87]) in connection with a special type of grid generator (cf. Roitzsch and Kornhuber [106]).

Similar situations frequently occur in other practical applications, motivating the adaptive resolution of the computational domain (cf. e.g. Kornhuber and Yserentant [88]) or the coarsening of a given fine mesh (cf. e.g. Bank and Xu [11], Chan and Smith [38], or Hackbusch and Sauter [67]). In this context, algebraic approaches also have become very popular. All these advanced multigrid techniques for selfadjoint linear problems can be combined with the nonlinear techniques developed above.

5.2.3 The A Posteriori Error Estimates

As in the previous example, we compare the *local* error estimate with the *iterative* error estimate generated by the truncated monotone multigrid

method with symmetric smoother. Again, the quality is measured by (approximations of) the efficiency indices κ_j (see p. 118). The results depicted in Figure 5.13 are better than one might have expected. Neither the (moderate) jump in the coefficients nor the aspect ratios have affected the performance of the error estimators. Moreover, the initial grid is fine enough to provide excellent estimates throughout the approximation. Again, we did not take advantage of the robustness of the iterative error estimate.

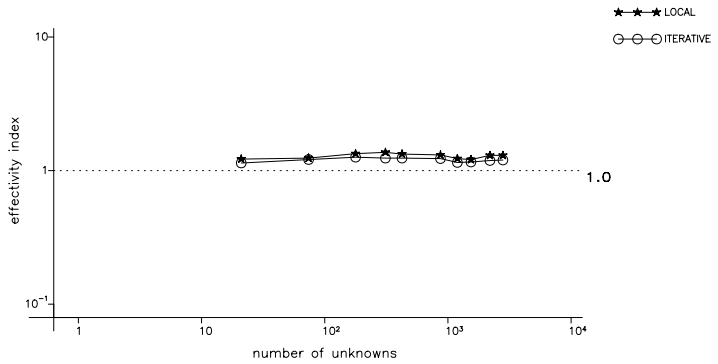

Figure 5.13 Effectivity indices of the approximation error estimates

In order to check the reliability of the stopping criterion for the iterative solution, we again consider the algebraic effectivity indices introduced in (1.3). As compared to the previous example, the slower convergence of the truncated monotone multigrid method leads to less accurate estimates of the algebraic error. The algebraic effectivity indices range from 0.37 to 0.55. However, on the last three levels, this underestimation is still compensated by the final correction step. We got similar results for the standard version.

5.3 Continuous Casting

Continuous casting is used in the steel industry for the rapid production of ingots. The essential features of the process are illustrated in Figure 5.14. The molten steel runs from a ladle into a water–cooled mold. After sufficient solidification at the surface, the product enters a secondary cooling region

where it is further cooled down by water sprays. After complete solidification, the ingot is cut off at a certain distance.

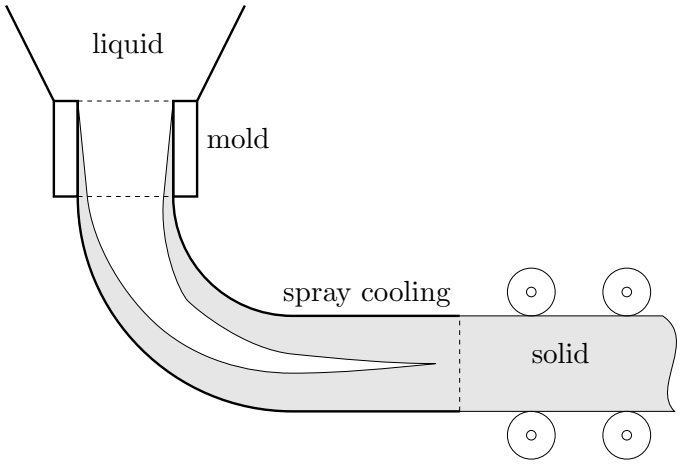

Figure 5.14 Schematic representation of continuous casting

Proper control of the cooling conditions in the mold and in the spray region is crucial for the whole process. Insufficient cooling may leave a liquid kernel at the cut–off point and cause a lot of damage. On the other hand, too much cooling may lead to cracks in the material and this may be even worse. Tests of suitable configurations on a production machine are extremely costly, motivating the numerical simulation of the whole process.

For given cooling conditions, we are interested in the stationary temperature distribution $\theta(x_1, x_2, x_3)$. The coordinates x_1, x_2 describe the *cross section* Ω of the bar and x_3 measures the distance from the beginning of the mold. Assuming *constant casting speed* v_{cast}, we introduce the Lagrange coordinates

$$x(t) = (x_1, x_2, x_3(t)), \quad x_3(t) = v_{cast}t.$$

Neglecting the heat conduction in the withdrawel direction x_3, the temperature $\theta = \theta(x(t), t)$ then satisfies the heat equation

$$\rho \frac{\partial}{\partial t}\mathcal{E}(\theta) + \nabla \cdot (\kappa(\theta)\nabla(\theta)) = 0 \qquad \text{in } \Omega \tag{3.9}$$

for all x_3. With reference to Section 1.1.2, ρ denotes the density, $\mathcal{E}$ is the specific internal energy, and κ is the thermal conductivity of the steel. Recall that $\mathcal{E}$ may be set–valued. In this formulation, the time $t = v_{cast}^{-1} x_3$ is nothing but a scaled space. Hence, the solution $\theta(x_1, x_2, t)$ of (3.9) can be interpreted as the desired stationary temperature distribution $\theta(x_1, x_2, v_{cast}^{-1} x_3)$ with respect to fixed spatial coordinates.

In the mold and in the spray region, heat is extracted by *convection and conduction*, giving rise to the Cauchy boundary conditions

$$-\kappa(\theta)\tfrac{\partial}{\partial n}\theta = q_{cool}(\theta - \theta_{cool}) \qquad \text{on } \partial\Omega, \tag{3.10}$$

where n denotes the outward normal on $\partial\Omega$. The heat transfer coefficient q_{cool} can be directly related to the rate of cooling water in the mold and in the sprays, respectively (cf. Laitinen and Neitaanmäki [89]) and θ_{cool} is the outward temperature. For the optimal control of q_{cool}, we refer to Neitaanmäki and Tiba [97] and literature cited therein. After the spray region, cooling takes place by *radiation* according to the Stefan–Boltzmann law. This leads to an additional *piecewise smooth* nonlinearity in the problem. We will not discuss this additional difficulty here, but refer to a related situation occurring in the final example.

Following Section 1.1.2, we now apply a standard Kirchhoff transformation $U = K(\theta)$ and a time discretization by the implicit Euler method to the parabolic problem (3.9). A weak formulation of the resulting spatial problems is given by

$$\begin{aligned} U_i \in H^1(\Omega) :\ & (W_{i-1}, v)_{L^2(\Omega)} - \tau_i (\nabla U_i, \nabla v)_{L^2(\Omega)} \in \\ & (\mathcal{H}(U_i), v)_{L^2(\Omega)} + \tau_i (q_{cool}(K^{-1}(U_i) - \theta_{cool}), v)_{L^2(\partial\Omega)},\ \forall v \in H^1(\Omega), \end{aligned}$$

where $\mathcal{H} = \rho\mathcal{E}(K^{-1}(\cdot))$ is the normalized enthalpy, $U_i \approx U(t_i)$ approximates the solution at time t_i, τ_i denotes the actual step size, and $W_{i-1} \in \mathcal{H}(U_{i-1})$ is a selection from the preceding time level.

As an outcome of the Cauchy boundary conditions, the spatial problems contain the additional nonlinearity $K^{-1}(U_i)$. Assuming that $\kappa(\theta)$ is piecewise constant, the inverse Kirchhoff transformation K^{-1} is a piecewise linear function. As the thermal conductivity is known to vary only moderately with θ, it is reasonable to use the explicit linearization

$$c(U_{i-1}) + b(U_{i-1})U_i \approx K^{-1}(U_i) = c(U_i) + b(U_i)U_i.$$

This leads to spatial problems of the form

$$u \in H: \quad \ell(v) - a(u,v) \in (\mathcal{H}(u), v)_{L^2(\Omega)}, \quad \forall v \in H, \tag{3.11}$$

where we have set $u = U_i$, $H = H^1(\Omega)$, and

$$a(v,w) = \tau_i(\nabla v, \nabla w)_{L^2(\Omega)} + \tau_i(q_{cool} b(U_{i-1})v, w)_{L^2(\partial\Omega)},$$
$$\ell(v) = (W_{i-1}, v)_{L^2(\Omega)} - \tau_i(q_{cool}(c(U_{i-1}) - \theta_{cool}), v)_{L^2(\partial\Omega)}.$$

Multigrid methods for the efficient solution of discrete analogues of (3.11) have been proposed by Hoppe and Kornhuber [73, 74] and Hoppe [72]. In their algorithms, the coarse grid correction is performed separately in each phase, leading to a poor coarse grid transport in comparison with truncated monotone multigrid methods (see Kornhuber [83]). Adaptive techniques for the two–phase Stefan problem have been developed by Nochetto, Paolini, and Verdi [98, 99]. Their local error indicators concentrate exclusively on the efficient resolution of the moving boundary. The resulting refinement strategy contains a considerable number of parameters.

Recall that (3.11) can be regarded as a special case of the reference problem (1.2.33). We emphasize that the completely implicit treatment of the nonlinearity $K^{-1}(U_i)$ leads to a straightforward extension of our approach to convex functionals of the form $\phi(v) = \int_\Omega \Phi_1(v)dx + \int_{\partial\Omega} \Phi_2(v)d\sigma$. This would also include problems of Signorini type.

In our numerical simulation, we consider a production machine with a mold of length 0.60 m, spray cooling takes place for further 6.00 m, and the cross section Ω of the ingot is shown in the left picture of Figure 5.15. The length of the horizontal edges of Ω is 0.1 m and the casting speed is $v_{cast} = 0.05\ m/s$.

We consider the production of steel with carbon content 0.12%. Following [117], the density is given by $\rho = 7.3 \cdot 10^3 kg/m^3$ and the volumetric enthalpy $\rho\mathcal{E}(\theta)$ is shown in the right picture of Figure 5.15. Observe that $\rho\mathcal{E}$ exhibits a large jump at the melting temperature $\theta_* = 1500\ ^oC$, but there are also small jumps occurring at 1470 ^oC and 1480 ^oC, respectively. This leads to a *multiphase Stefan problem*. For simplicity, we use the constant thermal conductivity $\kappa = 0.175\ kW/m\ ^oC$ in the liquidus $\theta > \theta_*$ and $\kappa = 0.05\ kW/m\ ^oC$ in the solidus $\theta < \theta_*$.

The cooling conditions are described by $q_{cool} = 1.5\ kW/m^2\ ^oC$, $\theta_{cool} = 80\ ^oC$ in the mold region and $q_{cool} = 1.0\ kW/m^2\ ^oC$, $\theta_{cool} = 27\ ^oC$ in the

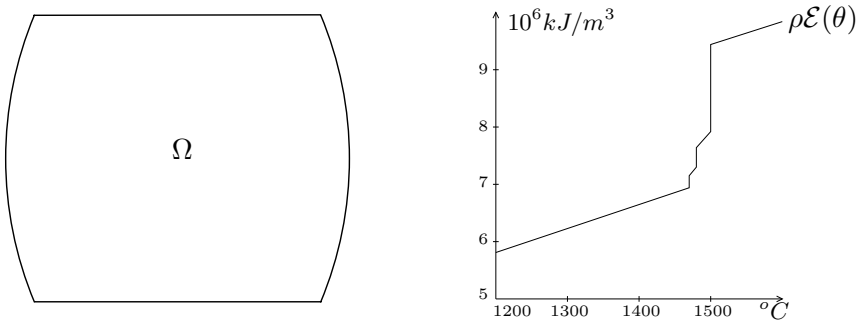

Figure 5.15 Cross section Ω and enthalpy $\rho\mathcal{E}$

spray region. Of course, one could also use more sophisticated functions q_{cool} (cf. e.g. Laitinen and Neitaanmäki [89]).

We start our computation at $t = 0$, i.e. at the beginning of the mold, assuming the constant initial temperature $\theta(x, 0) = 1501\ ^oC$. Our simulation ends at $T = 132s$, corresponding to the end of the spray region. We choose the step size $\tau_i = 1s$ in the mold and $\tau_i = 3s$ in the spray region, respectively. Of course, τ_i should be selected automatically based on a posteriori error estimates (see e.g. Bornemann [23, 24, 25]). This is a subject of current research.

5.3.1 The Adaptive Multilevel Method

The solution of the spatial problems (3.11) is carried out by the same adaptive algorithm as in the preceding examples, using the truncated monotone multigrid method (cf. Theorem 3.11) and the local a posteriori error estimate (cf. Theorem 4.3). Starting with the initial triangulation depicted in Figure 5.1, we successively approximate the curved boundary of Ω by moving the midpoints of all refined edges that lie on the approximating polygonal boundary to the exact boundary arc $\partial\Omega$. It is known for the linear self–adjoint case that such a modification does not affect the quality of the multigrid convergence rates (cf. Bramble and Pasciak [32]).

Collecting the results from all spatial problems, we can compose the three–dimensional stationary temperature distribution in the ingot. The profile along a diagonal is shown in Figure 5.16. The initial temperature of $1501\ ^oC$ is cooled down to $1203\ ^oC$ in the interior and to $512\ ^oC$ at the vertices.

Observe that complete solidification takes place just before the end of our simulation. Particularly in the spray region, a more accurate resolution of the solid/liquid interface can be obtained by smaller time steps.

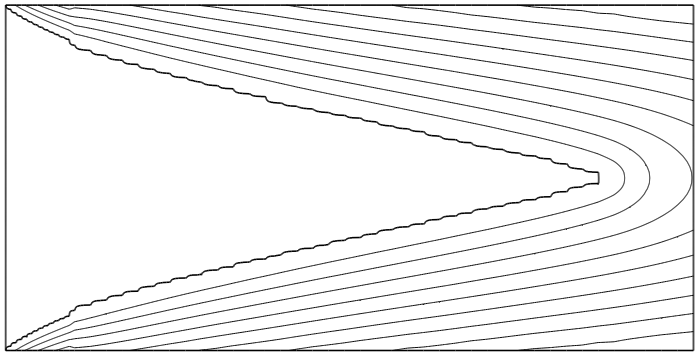

Figure 5.16 Stationary temperature distribution in the ingot

In order to illustrate the behavior of our adaptive algorithm, we consider the time level $t_{12} = 12s$ (at the end of the mold) and the last time level $t_{52} = 132s$ (at the end of the spray region) in more detail. The resulting final triangulations $\mathcal{T}_{16}$ and $\mathcal{T}_8$ together with the level curves of the corresponding final approximations $\tilde{u}_{16}$ and $\tilde{u}_8$ are depicted in Figure 5.17 and Figure 5.18, respectively. Recall that, in this chapter, the subscripts denote the number of refinement steps.

At the end of the mold, the width of the solid region is almost $1cm$. The refinement concentrates on the three interfaces, reflecting the jumps of $\rho\mathcal{E}$, and on the boundary $\partial\Omega$, where the cooling takes place. These are the two essential features of the whole process. In the liquid region, the approximate solution is almost constant. Hence, there is as little refinement as possible.

At the end of the spray region, the solid region covers the whole cross section Ω so that the last spatial problem reduces to a *linear* variational equality. The adaptive algorithm produces an almost uniformly refined final triangulation $\mathcal{T}_9$.

We now compare the approximation histories shown in Table 5.3 and Table 5.4, respectively. At first sight, we observe a much larger number of refinement steps in the nonlinear case. One could easily reduce this additional computational effort for the adaptive location of the interfaces by simple heuristics. We will come back to this point in the final example.

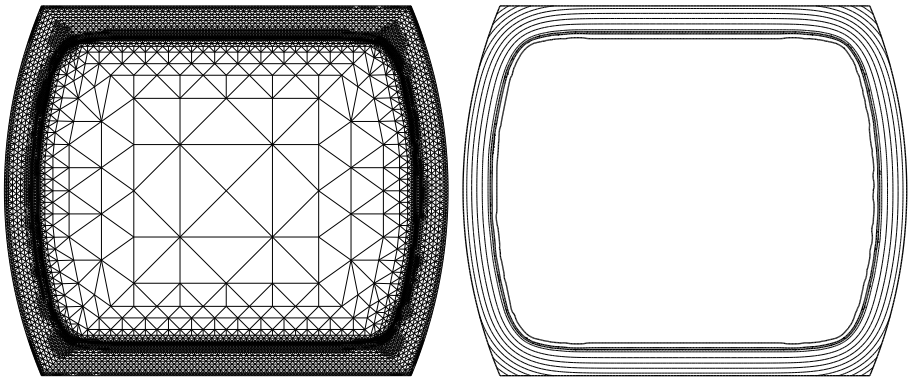

Figure 5.17 Final triangulation $\mathcal{T}_{16}$ and final approximation $\tilde{u}_{16}$ at the end of the mold

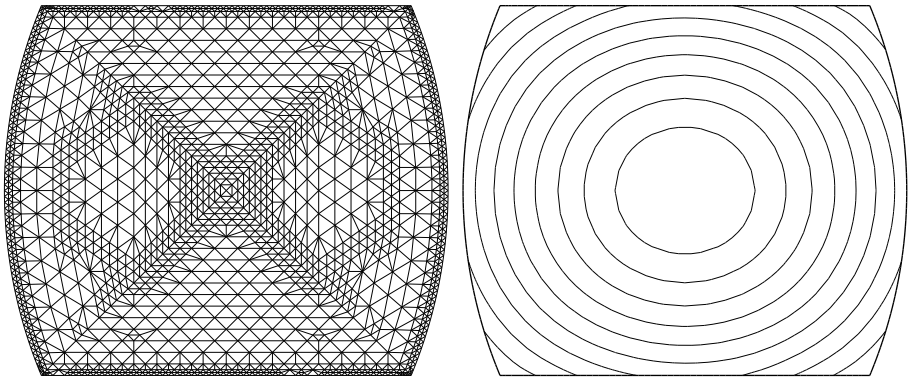

Figure 5.18 Final triangulation $\mathcal{T}_8$ and final approximation $\tilde{u}_8$ at the end of the sprays

The spatial problems (3.11) are singularly perturbed with respect to the time step size τ_i so that it may be dangerous to measure the spatial error in the energy norm. This explains the severe overestimation of the true approximation error on the first refinement levels. Following Bornemann [23, 24, 25] it is more appropriate to measure the error in a scaled norm of the form $\|\!| v |\!\| = (\frac{1}{1+\tau}(v,v)_{L^2(\Omega)} + \frac{\tau}{1+\tau}a(v,v))^{1/2}$ which makes sense for $\tau \to 0$ and $\tau \to \infty$. A posteriori error estimates in such norms follow directly from the results in Chapter 4.

Level	Depth	Nodes	Iterations	Error %
0	0	5	2	16577.7
1	1	13	2	6455.0
2	2	37	2	1374.1
3	3	121	2	349.1
4	4	305	2	117.1
5	5	373	2	81.4
6	6	468	3	62.8
7	6	879	3	42.0
8	6	1077	3	30.7
9	7	1416	3	22.2
10	7	1798	3	17.0
11	7	2493	3	13.3
12	8	3149	2	10.0
13	8	4141	2	8.3
14	8	5857	2	6.7
15	9	8313	2	5.5
16	9	10185	2	4.7

Table 5.3 Approximation History at the end of the mold

Level	Depth	Nodes	Iterations	Error %
0	0	5	2	1503.3
1	1	13	2	412.3
2	2	37	2	102.5
3	3	125	2	29.5
4	4	337	2	13.4
5	5	709	2	8.2
6	6	873	2	6.6
7	7	1311	2	5.4
8	7	1689	2	4.6

Table 5.4 Approximation History at the end of the sprays

As in the previous example, we observe an asymptotic error reduction of at least $\mathcal{O}(n_j^{-1/2})$ for both spatial problems. There is also not much difference concerning the number of iterations, which can hardly be reduced.

5.3.2 The Monotone Multigrid Methods

In order to compare the standard and the truncated version of our monotone multigrid methods, we consider the discrete spatial problem arising on time level $t_{12} = 12s$ (at the end of the mold) in more detail.

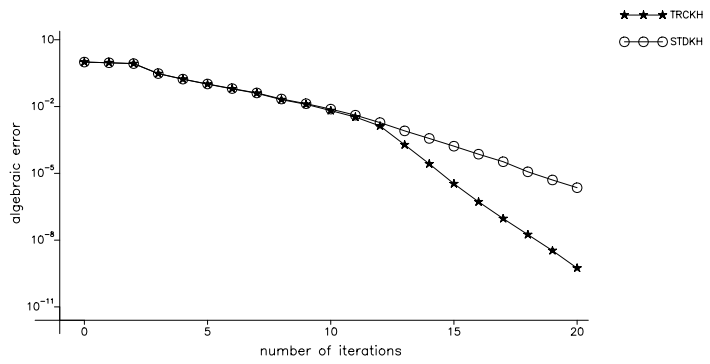

Figure 5.19 Iteration history: Initial iterate $u_j^0 = 0$

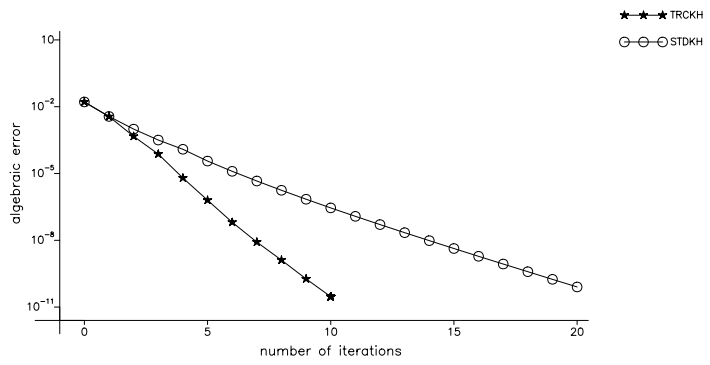

Figure 5.20 Iteration history: Interpolated initial iterate

As in the preceding examples, we first apply both methods to the discrete problem on the final refinement level $j = 16$, choosing the initial iterate $u_j^0 = 0$. The resulting iteration history, depicted in Figure 5.19, reminds us

of the related experiment for the membrane problem (cf. Figure 5.3). As usual, we observe a leading transient phase reflecting the nonlinearity of the problem, followed by an asymptotic phase corresponding to the reduced linear case. Again, the asymptotic convergence rate of the standard version (STDKH) suffers from poor coarse grid transport as compared with the truncated monotone multigrid method (TRCKH).

Starting with the interpolated solution from the previous level, we obtain the iteration history illustrated in Figure 5.20. In this case, the transient phase is completely skipped.

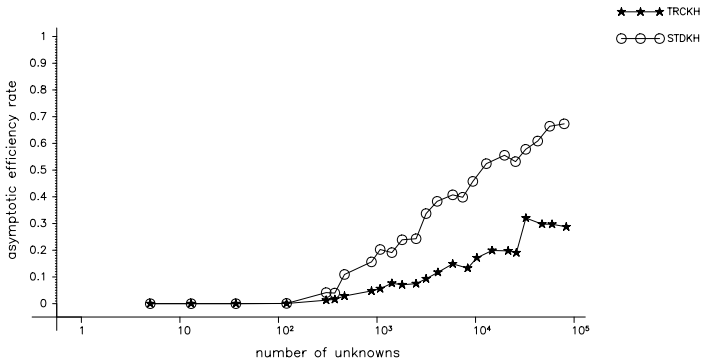

Figure 5.21 Asymptotic efficiency rates

To illustrate the convergence behavior for increasing refinement, we consider the asymptotic efficiency rates ρ_j, $j = 1, \ldots, 16$, defined according to (1.1).

	$t_1 = 1s$	$t_6 = 6s$	$t_{12} = 12s$	$t_{32} = 72s$	$t_{52} = 132s$
TRCKH	0.19	0.20	0.19	0.11	0.08
STDKH	0.45	0.53	0.51	0.32	0.08

Table 5.5 Asymptotic Efficiency rates on various time levels

The results are shown in Figure 5.21. The efficiency rates of STDKH and TRCKH are the same for $j = 0, 1, 2, 3$, because on these grids the solid region is not yet resolved. Then, STDKH exhibits a slow but steady increase of ρ_j while for TRCKH the efficiency rates seem to saturate at about 0.3. Table 5.5 displays the asymptotic efficiency rates for various time levels t_k

and the corresponding final triangulations. The number of nodes decreases from almost $3 \cdot 10^4$ for $t_1 = 1s$ to about $4 \cdot 10^3$ on the last time level. For the remaining spatial problems, we obtained similar results.

5.3.3 The A Posteriori Error Estimates

Again, we consider the spatial problem arising at $t_{12} = 12s$. This time, we compare the local error estimate (cf. Theorem 4.3) with the iterative error estimate generated by the truncated monotone multigrid method with symmetric smoother (cf. Theorem 4.5).

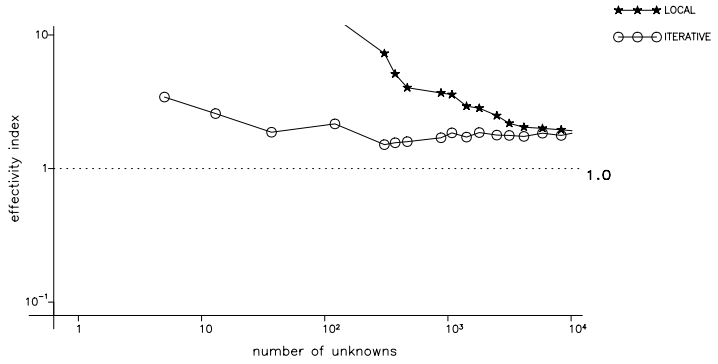

Figure 5.22 Effectivity indices of the approximation error estimates

The corresponding (approximate) effectivity indices are computed in the same way as before (see p. 118) and the results are shown in Figure 5.22. As already mentioned above, the local error estimator considerably overestimates the true error on the leading coarse triangulations. However, we obtain much better results on higher levels. The second error estimator based on nonlinear iteration is much more robust. This approach benefits from the more careful decoupling of the discrete defect problems using also the off–diagonal elements of the stiffness matrix. We observed very similar results on the other time levels, particularly for the linear problem at $t_{52} = 132s$.

As a consequence of the very fast convergence of the monotone multigrid methods, we found very satisfying algebraic effectivity rates, ranging from 0.79 to 1.0 for the truncated monotone multigrid method and from 0.67 to 1.0 for the standard variant.

5.4 The Porous Medium Equation

In our final example, we consider the degenerate parabolic differential equation

$$\frac{\partial}{\partial t}\rho = \Delta(\beta\rho_+^m) \qquad \text{in } \Omega, \tag{4.12}$$

describing the adiabatic flow of a homogeneous gas with density ρ through a porous medium (cf. Section 1.1.3). Elliptic–parabolic problems arising for example in nonstationary filtration can be treated in a similar way.

As outlined in Section 1.1.3, we first use the Kirchhoff-type transformation $\rho = \mathcal{P}(U)$ to reformulate (4.12) as the differential inclusion (1.1.30) which is then discretized by an implicit Euler scheme providing the spatial problems

$$u \in H: \quad \ell(v) - a(u,v) \in (\mathcal{P}(u), v)_{L^2(\Omega)}, \quad \forall v \in H, \tag{4.13}$$

where $u = U_i$ approximates the solution on the time level t_i, the solution space H incorporates the essential boundary conditions, and we have set

$$a(v,w) = \tau_i \int_\Omega \nabla v \cdot \nabla w \, dx, \quad \ell(v) = \int_\Omega W_{i-1} v \, dx$$

with τ_i denoting the time step size and W_{i-1} being a selection of the normalized enthalpy from the preceding time level.

Choosing Φ in such a way that $\partial\Phi = \mathcal{P}$, we have to solve a minimization problem of the form (1.2.33) in each time step. However, in contrast to the preceding examples, such a function Φ is *not piecewise quadratic* so that we cannot apply our monotone multigrid methods directly. One possible remedy is the *piecewise linear approximation* $\mathcal{P}_M$ of $\mathcal{P}$ which should be performed in such a way that the finite speed of the free boundary is preserved.

To derive such an approximation $\mathcal{P}_M$, we introduce the grid points $\theta_i = ((i-1)/M)^m$, $i = 1, \dots, M$. For $z \geq \theta_2$ the values $\mathcal{P}_M(z)$ are obtained by piecewise linear interpolation of $\mathcal{P}(\theta_i)$, $i = 2, \dots, M$. In the initial interval (θ_1, θ_2), we use the linear extrapolation from (θ_2, θ_3), giving a positive value $H_1 > 0$ at $\theta_1 = 0$ as illustrated in Figure 5.23. This leads to the approximation $\mathcal{P}_M(0) = [H_1, -\infty) \supset \mathcal{P}(0) = [0, \infty)$ at the singularity $z = 0$, which is crucial for the finite speed of the discrete free boundary.

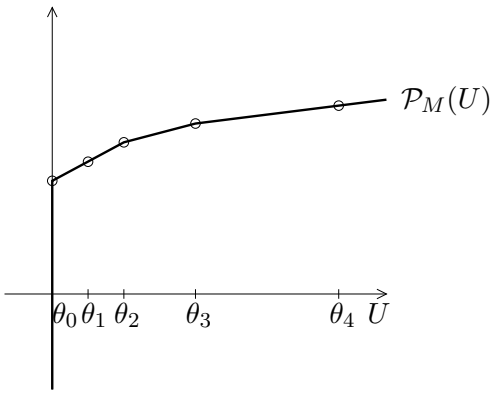

Figure 5.23 Approximation $\mathcal{P}_M$ of $\mathcal{P}$

The accuracy of this approximation was checked numerically for a two–dimensional analogue of the Barenblatt solution (cf. e.g. Jäger and Kačur [77]). Moreover, in most situations of practical interest, the crucial function $\mathcal{P}$ itself can be determined only approximately, for example by the fitting of experimental data. Hence, using $\mathcal{P}_M$ instead of $\mathcal{P}$ can be justified from a practical point of view. Nevertheless, a tempting alternative to this a priori linearization is to extend monotone multigrid methods to functionals ϕ which are not piecewise quadratic but *piecewise smooth*. This will be the subject of a forthcoming paper.

Other discretizations of (4.12) were studied by Berger, Brézis and Rogers [21], Magenes, Nochetto and Verdi [91], and Nochetto and Verdi [100]. Such linear schemes do not preserve the finite speed of propagation of perturbations. As a consequence, sharp layers occurring in the exact solution may be smeared out in course of the approximation. This problem was remedied by Jäger and Kačur [77] at the expense of an additional fixed point iteration for each nodal point. We refer to Bänsch [14] for numerical experiments. Damped Newton–like linearizations in connection with regularization techniques and a multigrid solution of the linear sub–problems have been proposed by Fuhrmann [59, 60].

In our numerical experiment, we choose $m = 2$, $\beta = 1$, and the initial condition

$$\rho_0(x_1, x_2) = \Big[0.4 - r^2(1 + 0.5\sin(14\varphi))\Big]_+^{1/2}$$

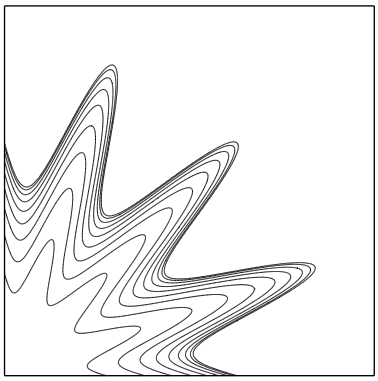

Figure 5.24 Initial density distribution ρ_0

with $r = (x_1^2 + x_2^2)^{1/2}$ and $\varphi = \text{atan}(x_2/x_1)$. The level curves of ρ_0 are shown in Figure 5.24. We consider the spatial domain $\Omega = (0,1) \times (0,1)$ and the time interval $(0,T) = (0, 0.05)$, using the uniform time step size $\tau = 1/800$. Homogeneous Dirichlet conditions are prescribed at $x_1 = 1$, $x_2 = 1$ and we impose homogeneous Neumann conditions at $x_1 = 0$, $x_2 = 0$. For the approximation $\mathcal{P}_M$ of $\mathcal{P}$, we choose $M = 100$.

5.4.1 The Adaptive Multilevel Method

The spatial problems (4.13), with $\mathcal{P}$ replaced by $\mathcal{P}_M$, are now solved approximately by our adaptive algorithm. Due to the a priori linearization, these problems involve *a very large number of phases.*

As the optimal resolution of the free boundary is of prominent importance for the proper approximation of the density ρ, we additionally refine all triangles satisfying $0 < \tilde{u}_j(p) < 10^{-5}$ for at least one vertex p. As our discretization provides discrete solutions with compact support, there will be no refinement in the unsaturated phase. In order to limit the computational costs, we prescribe the maximal refinement depth $d_{max} = 8$.

Apart from these modifications, we use the same adaptive multilevel algorithm as in the preceding examples incorporating the truncated monotone multigrid method and the local a posteriori error estimate. The initial triangulation $\mathcal{T}_0$ is taken from our first example (see Figure 5.1).

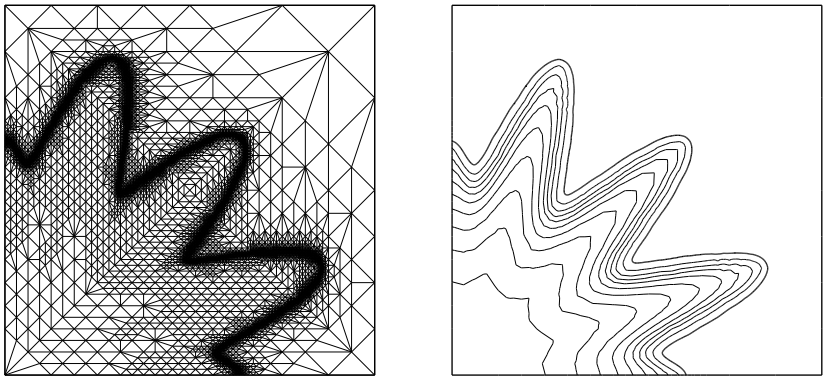

Figure 5.25 Final triangulation $\mathcal{T}_8$ and final approximation $\tilde{u}_8$ for the first time step

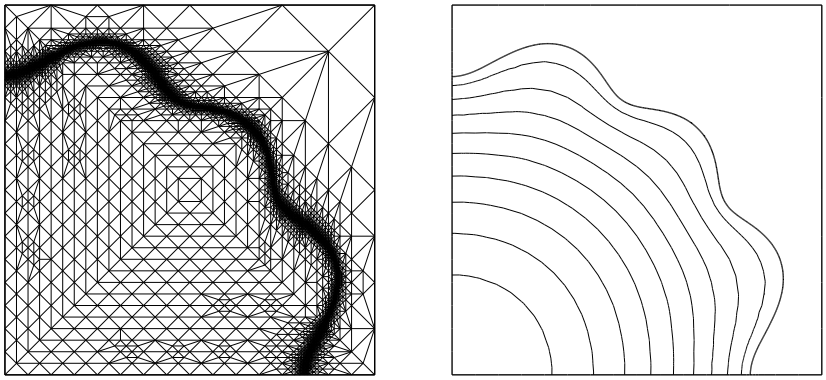

Figure 5.26 Final triangulation $\mathcal{T}_8$ and final approximation $\tilde{u}_8$ for the last time step

In order to illustrate the behavior of the adaptive algorithm, we consider the first lime level t_1 and the last time level t_{40} in some detail. The final triangulations and the corresponding final approximations are shown in Figures 5.25 and 5.26, respectively. As intended, the nodal points are concentrated in the neighborhood of the approximate free boundary. The triangulations reflect the smoothness of the solution in the saturated phase. Both grids are as coarse as possible where the density is zero. Observe that the discretization preserves steep gradients at the interface while the interface itself is smoothed.

The corresponding approximation histories are shown in Table 5.6 and Table 5.7, respectively. Recall from the previous example that the semi–discrete problems (4.13) may become singularly perturbed for small time steps. Hence, it is not astonishing that the approximation error is again heavily overestimated on the first levels. This motivates error estimates in other norms such as the L^2-norm or the scaled norm introduced above (cf. p. 133). As adaptive refinement is now based on local error indicators *and* additional heuristic criteria, we can no longer expect an $\mathcal{O}(n_j)$ behavior of the approximation error. In fact, the error is reduced very slowly with increasing refine-

Level	Depth	Nodes	Iterations	Error %
0	0	2	1	11225.0
1	1	8	2	8660.3
2	2	29	3	1542.7
3	3	104	3	206.4
4	4	349	3	77.8
5	5	850	4	36.1
6	6	1976	4	27.1
7	7	3996	3	25.7
8	8	8143	3	22.9

Table 5.6 Approximation history for the first time step

Level	Depth	Nodes	Iterations	Error %
0	0	2	2	14071.2
1	1	8	2	2687.0
2	2	32	2	371.4
3	3	116	2	66.8
4	4	298	2	66.2
5	5	660	2	48.4
6	6	1325	2	32.2
7	7	2555	2	24.0
8	8	5113	2	19.8

Table 5.7 Approximation history for the last time step

ment so that the prescribed accuracy of 5.% is not reached within the maximal number of $d_{max} = 8$ steps. In spite of the strong nonlinearity of the problem, the number of iterations still remains very moderate. This illustrates the robustness of our multigrid approach.

5.4.2 The Monotone Multigrid Methods

We perform the same experiments as above (with very similar results). Let

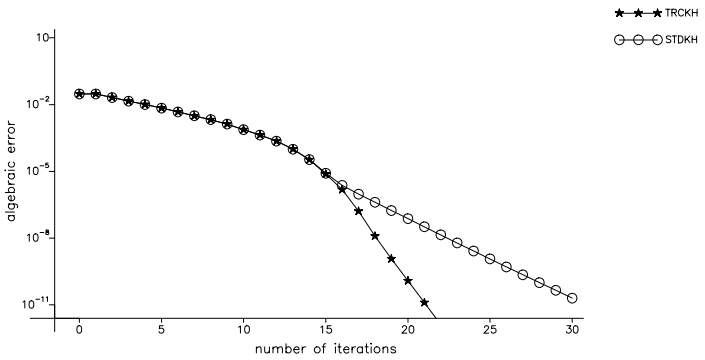

Figure 5.27 Iteration history: Initial iterate $u_j^0 = 0$

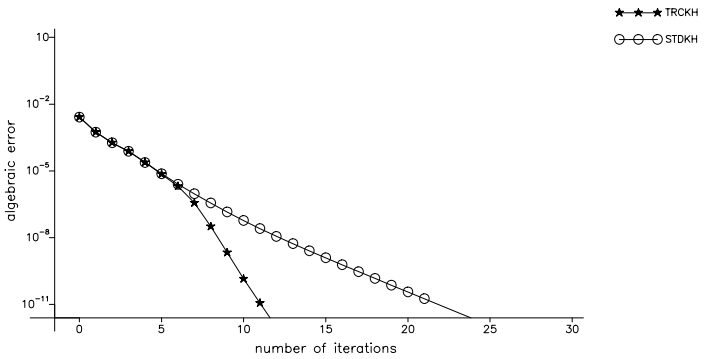

Figure 5.28 Iteration history: Interpolated initial iterate

us consider the discrete spatial problem arising in the first time step on the final refinement level $j = 8$. Starting with the initial iterate $u_j^0 = 0$, we obtain the iteration history depicted in Figure 5.27. As compared to the Stefan problem, the much larger number of phases causes a larger number of leading transient steps but does not affect the rapid asymptotic convergence of the truncated version. Figure 5.28 shows the performance of the multigrid

algorithms in case of the interpolated initial iterate. In contrast to the preceding examples, the transient phase does not vanish completely. However, even the transient convergence is considerably faster taking advantage of the fast reduction of the high frequencies and of the small time step. For larger time steps, we observed similar asymptotic convergence rates as in the previous example.

To illustrate the convergence behavior for increasing j, we computed the asymptotic efficiency rates ρ_j according to (1.2). The results are shown in Figure 5.29. The increasing efficiency rates of the truncated version reflect the increasing number of transient iteration steps. Again, this can be explained by the large number of phases. Table 5.8 displays the asymptotic

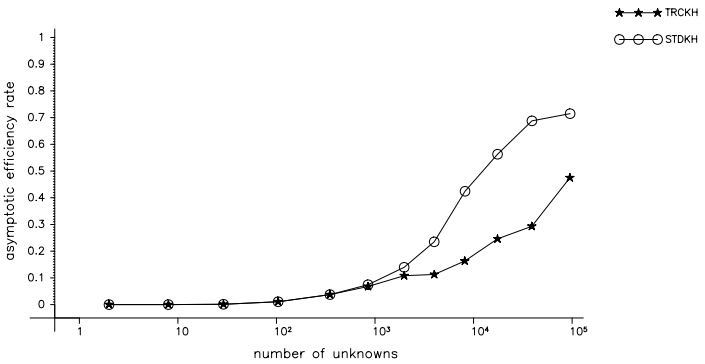

Figure 5.29 Asymptotic efficiency rates

efficiency rates for other times t_k and the final refinement level $j = 8$. Similar results were obtained for the remaining time steps.

	$t_8 = 0.01$	$t_{16} = 0.02$	$t_{24} = 0.03$	$t_{32} = 0.04$	$t_{40} = 0.05$
TRCKH	0.10	0.14	0.11	0.08	0.09
STDKH	0.43	0.37	0.35	0.34	0.35

Table 5.8 Asymptotic efficiency rates for various time steps

5.4.3 The A Posteriori Error Estimates

We briefly discuss the behavior of the error estimates for the first time step which again is typical for the remaining spatial problems. The (approxi-

mate) effectivity indices (cf. p. 118) of our local and iterative estimates of the approximation error are shown in Figure 5.30. As already mentioned

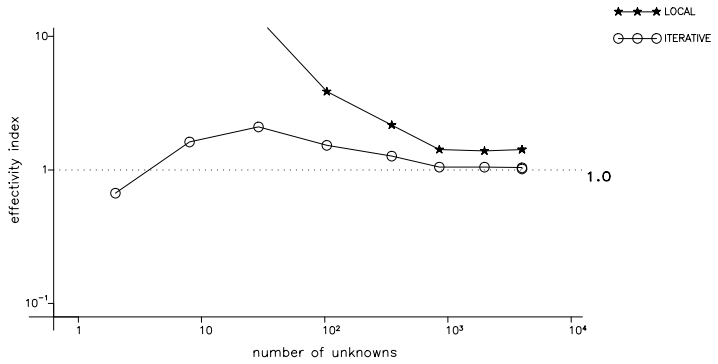

Figure 5.30 Approximation error estimates

above, the local error estimate should be used carefully on coarse triangulations, but has very satisfying asymptotic properties. Again, the (more expensive) iterative error estimate provides much better results throughout the approximation.

For the truncated monotone multigrid method, the algebraic effectivity indices are varying from 0.68 to 1.1 and we observed similar results for the standard version.

Bibliography

[1] R.A. Adams. *Sobolev Spaces*. Academic Press, New York, 1975.

[2] H.W. Alt and S. Luckhaus. Quasilinear elliptic–parabolic differential equations. *Math. Z.*, 183:311–341, 1983.

[3] D.G. Aronson. The porous medium equation. In A. Fasano and M. Primicerio, editors, *Nonlinear Diffusion Problems*, volume 1224 of *Lecture Notes in Mathematics*, pages 1–46, New York, 1986. Springer.

[4] J.P. Aubin. *Optima and Equilibria*. Springer, Berlin, 1993.

[5] W. Aunziger and H.J. Stetter. Defect corrections and multigrid iterations. In *Multigrid Methods*, volume 960 of *Lecture Notes in Mathematics*, pages 327–351, Berlin, 1982. Springer.

[6] I. Babuška and W.C. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.*, 15:736–754, 1978.

[7] C. Baiocchi and A. Capelo. *Variational and Quasivariational Inequalities*. Wiley, Chichester, 1984.

[8] R.E. Bank. *PLTMG – A Software Package for Solving Elliptic Partial Differential Equations, User's Guide 6.0*. Frontiers in Applied Mathematics. SIAM, Philadelphia, 1990.

[9] R.E. Bank, T.F. Dupont and H. Yserentant. The hierarchical basis multigrid method. *Numer. Math.*, 52:387–404, 1988.

[10] R.E. Bank and R.K. Smith. A posteriori error estimates based on hierarchical bases. *SIAM J. Num. Anal.*, 30:921–935, 1993.

[11] R.E. Bank and J. Xu. An algorithm for coarsening unstructured meshes. *Preprint*, 1994.

[12] R.E. Bank and J. Xu. The hierarchical basis multigrid method and incomplete LU decomposition. In Keyes and Xu [80], pages 163–174.

[13] E. Bänsch. Local mesh refinement in 2 and 3 dimensions. *IMPACT Comput. Sci. Engrg.*, 3:181–191, 1991.

[14] E. Bänsch. Numerical experiments with adaptivity for the porous media equation. To appear in AMUC.

[15] V. Barbu. *Nonlinear Semigroups and Differential Equations in Banach Spaces*. Noordhoff International Publishing, Leyden, 1976.

[16] C Bäsler and W. Törnig. On monotone including nonlinear multigrid methods and applications. *Computing*, 50:51–67, 1993.

[17] P. Bastian. *Adaptive parallele Mehrgitterverfahren*. Skripten zur Numerik. Teubner, Stuttgart, 1995.

[18] J. Bear. *Dynamics of Fluids in Porous Media*. Elsevier, New York, 1972.

[19] R. Beck, B. Erdmann and R. Roitzsch. KASKADE Manual, Version 3.0. Technical Report TR95-4, Konrad–Zuse–Zentrum (ZIB), Berlin, 1995.

[20] R. Becker, C. Johnson and R. Rannacher. Adaptive error control for multigrid finite element methods. *Computing*, 55:271–288, 1995.

[21] A.E. Berger, H. Brezis and J.C.W. Rogers. A numerical method for solving the problem $u_t - \delta f(u) = 0$. *R.A.I.R.O. Analyse numérique*, 13:297–312, 1979.

[22] J. Bey. Analyse und Simulation eines Konjugierte-Gradienten-Verfahrens mit einem Multilevel-Präkonditionierer zur Lösung dreidimensionaler, elliptischer Randwertprobleme für massiv parallele Rechner. Master's thesis, RWTH Aachen, 1991.

[23] F.A. Bornemann. An adaptive multilevel approach to parabolic equations I. General theory and 1D-implementation. *IMPACT Comput. Sci. Engrg.*, 2:279–317, 1990.

[24] F.A. Bornemann. An adaptive multilevel approach to parabolic equations II. Variable-order time discretization based on a multiplicative error correction. *IMPACT Comput. Sci. Engrg.*, 3:93–122, 1991.

[25] F.A. Bornemann. An adaptive multilevel approach to parabolic equations III. 2D Error estimation and multilevel preconditioning. *IMPACT Comput. Sci. Engrg.*, 4:1–45, 1992.

[26] F.A. Bornemann and P. Deuflhard. Cascadic multigrid methods for elliptic problems. *Numer. Math.*, to appear.

[27] F.A. Bornemann, B. Erdmann and R. Kornhuber. Adaptive multilevel methods in three space dimensions. *Int. J. Numer. Meth. Engrg.*, 36:3187–3203, 1993.

[28] F.A. Bornemann, B. Erdmann and R. Kornhuber. A posteriori error estimates for elliptic problems in two and three space dimensions. *SIAM J. Numer. Anal.*, 33:1188–1204, 1996.

[29] F.A. Bornemann and H. Yserentant. A basic norm equivalence in the theory of multilevel methods. *Numer. Math.*, 64:455–476, 1993.

[30] D. Braess. *Finite Elemente*. Springer, Berlin, 1991.

[31] J.H. Bramble. *Multigrid Methods*. Pitman Research Notes in Mathematics. Longman, Harlow, 1993.

[32] J.H. Bramble and J.E. Pasciak. New estimates for multilevel algorithms including the V–cycle. *Math. Comp.*, 60:447–471, 1993.

[33] A. Brandt and C.W. Cryer. Multigrid algorithms for the solution of linear complementary problems arising from free boundary problems. *SIAM J. Sci. Stat. Comput.*, 4:655–684, 1983.

[34] H. Brézis. *Monotonicity Methods in Hilbert Spaces and some Applications to Partial Differential Equations*. Academic Press, New York, 1971.

[35] H. Brézis. *Opérateurs Maximaux Monotones et Semi-groupes de Contractions dans les Espaces de Hilbert*. North–Holland, New York, 1973.

[36] F. Brezzi, W.W. Hager and P.A. Raviart. Error estimates for the finite element solution of variational inequalities I. *Numer. Math.*, 28:431–443, 1977.

[37] L.A. Caffarelli and A. Friedman. Regularity of the free boundary for a gas in an n-dimensional porous medium. *Indiana Univ. Math. J.*, 29:361–391, 1980.

[38] T.F. Chan and B. Smith. Domain decomposition and multigrid algorithms for elliptic problems on unstructured meshes. In Keyes and Xu [80], pages 175–189.

[39] P.G. Ciarlet. Basic error estimates for elliptic problems. In *Finite Element Methods (Part 1)*, volume II of *Handbook of Numerical Analysis*, pages 17–351, Amsterdam, 1991. North–Holland.

[40] F.H. Clarke. *Optimization and Nonsmooth Analysis*. John Wiley and Sons, New York, 1983.

[41] J. Crank. *Free and Moving Boundary Problems*. Oxford University Press, Oxford, 1988.

[42] H. Darcy. *Les fontaines publiques de la ville de Dijon*. Dalmont, Paris, 1856.

[43] K. Deimling. *Nonlinear Analysis*. Springer, Berlin, 1985.

[44] P. Deuflhard. Cascadic conjugate gradient methods for elliptic partial differential equations. Algorithm and results. In Keyes and Xu [80], pages 29–42.

[45] P. Deuflhard, P. Leinen and H. Yserentant. Concepts of an adaptive hierarchical finite element code. *IMPACT Comput. Sci. Engrg.*, 1:3–35, 1989.

[46] W. Dörfler. A convergent adaptive algorithm for Poisson's equation. To appear in SIAM J. Numer. Anal., 1996.

[47] M. Dryja and O.B. Widlund. Multilevel Additive methods for elliptic finite element problems. In W. Hackbusch, editor, *Parallel Algorithms for Partial Differential Equations*, Braunschweig, 1991. Vieweg.

[48] G. Duvaut and J.L. Lions. *Les inéquations en mécanique et en physique*. Dunaud, Paris, 1972.

[49] I. Ekeland and R. Temam. *Convex Analysis and Variational Problems*. North–Holland, Amsterdam, 1976.

[50] C.M. Elliott. On the finite element approximation of an elliptic variational inequality arising from an implicit time discretization of the Stefan problem. *IMA J. Numer. Anal.*, 1:115–125, 1981.

[51] C.M. Elliott and J.R. Ockendon. *Weak and Variational Methods for Moving Boundary Problems*, volume 53 of *Research Notes in Mathematics*. Pitman, London, 1982.

[52] B. Erdmann, M. Frei, R.H.W. Hoppe, R. Kornhuber and U. Wiest. Adaptive finite element methods for variational inequalities. *East-West J. Numer. Math.*, 1(3):165–197, 1993.

[53] B. Erdmann, R.H.W. Hoppe and R. Kornhuber. Adaptive multilevel methods for obstacle problems in three space dimensions. In W. Hackbusch and G. Wittum, editors, *Adaptive Methods – Algorithms, Theory and Computations*, pages 120–142, Braunschweig, 1994. Vieweg.

[54] W. Feiler and W. Gerlach. The reverse voltage limit of planar *pn*–junctions with field plates. *Archiv f. Elektrotechnik*, 75:223–229, 1992.

[55] G. Fichera. Problemi elastotatici con vincoli unilaterali: il problema di Signorini con ambigue condizioni al contorno. *Atti. Acad. Naz. Lincei Mem. Cl. Sci. Fis. Mat. Nat. Sez. Ia*, 7(8):91–140, 1963-1964.

[56] A. Friedman. The Stefan problem in several space variables. *Trans. Amer. Math. Soc.*, 133:51–87, 1968.

[57] A. Friedman. *Variational Principles and Free Boundary Problems*. Wiley, New York, 1982.

[58] A. Friedman and J. Spruck. *Variational and Free Boundary Problems*. Springer, New York, 1993.

[59] J. Fuhrmann. Calculation of saturated – unsaturated flow with a Newton–multigrid method. In S. Hengst, editor, *GAMM–Seminar on Multigrid–Methods*, pages 2–14, Berlin, 1993. IAAS Report No. 5.

[60] J. Fuhrmann. On the numerical solution of the equation of saturated/unsaturated flow in porous media. In A. Peters, G. Wittum, B. Herrling, U. Meissner, C.A. Brebbia, W.G. Gray and G.F. Pinder, editors, *Computational Methods in Water Resources X*, pages 1393–1400, Dordrecht, 1994. Kluwer Academic Publishers.

[61] R. Glowinski. *Numerical Methods for Nonlinear Variational Problems*. Springer, New York, 1984.

[62] R. Glowinski, J.L. Lions and Trémolières. *Numerical Analysis of Variational Inequalities*. North–Holland, Amsterdam, 1981.

[63] M. Griebel. *Multilevelmethoden als Iterationsverfahren über Erzeugendensystemen*. Skripten zur Numerik. Teubner, Stuttgart, 1994.

[64] M. Griebel and P. Oswald. On the abstract theory of additive and multiplicative Schwarz algorithms. *Numer. Math.*, 70:163–180, 1995.

[65] W. Hackbusch. *Multi-Grid Methods and Applications*. Springer, Berlin, 1985.

[66] W. Hackbusch and H.D. Mittelmann. On multigrid methods for variational inequalities. *Numer. Math.*, 42:65–76, 1983.

[67] W. Hackbusch and S.A. Sauter. Composite finite elements for the approximation of pdes on domains with complicated micro–structures. Bericht 9505, Universität Kiel, 1995.

[68] I. Hlaváček, J. Haslinger, J. Nečas and J. Lovíšek. *Solution of Variational Inequalities in Mechanics*. Springer, Berlin, 1988.

[69] R.H.W. Hoppe. Multigrid algorithms for variational inequalities. *SIAM J. Numer. Anal.*, 24:1046–1065, 1987.

[70] R.H.W. Hoppe. Two-sided approximations for unilateral variational inequalities by multigrid methods. *Optimization*, 18:867–881, 1987.

[71] R.H.W. Hoppe. Une méthode multigrille pour la solution des problèmes d'obstacle. M^2 *AN*, 24:711–736, 1990.

[72] R.H.W. Hoppe. A globally convergent multigrid algorithm for moving boundary problems of two-phase Stefan type. *IMA J. Numer. Anal.*, 13:235–253, 1993.

[73] R.H.W. Hoppe and R. Kornhuber. Multigrid methods for the two-phase Stefan problem. In S.F. McCormick, editor, *Multigrid Methods*, pages 267–297, New York, 1988. Marcel Dekker.

[74] R.H.W. Hoppe and R. Kornhuber. Multigrid solution of two coupled Stefan equations arising in induction heating of large steel slabs. *Int. J. Numer. Meth. Engrg.* 30:779–801, 1990.

[75] R.H.W. Hoppe and R. Kornhuber. Adaptive multilevel–methods for obstacle problems. *SIAM J. Numer. Anal.*, 31(2):301–323, 1994.

[76] C. Hunt and N. Nassif. On a variational inequality and its application in the theory of semiconductors. *SIAM J. Numer. Anal.*, 12:938–950, 1975.

[77] W. Jäger and J. Kačur. Solution of porous medium type systems by linear approximation schemes. *Numer. Math.*, 60:407–427, 1991.

[78] J.W. Jerome. *Approximation of Nonlinear Evolution Equations*. Academic Press, New York, 1983.

[79] S. Kamenomostskaya. On the Stefan problem. *Mat. Sb.*, 53:489–514, 1961.

[80] D.E. Keyes and J. Xu, editors. *Proceedings of the 7th International Conference on Domain Decomposition Methods 1993.*, Providence, 1994. AMS.

[81] D. Kinderlehrer and G. Stampacchia. *An Introduction to Variational Inequalities and Their Applications*. Academic Press, New York, 1980.

[82] R. Kornhuber. Monotone multigrid methods for elliptic variational inequalities I. *Numer. Math.*, 69:167 – 184, 1994.

[83] R. Kornhuber. Monotone multigrid methods for elliptic variational inequalities II. *Numer. Math.*, 72:481 – 499, 1996.

[84] R. Kornhuber. A posteriori error estimates for elliptic variational inequalities. *Computers Math. Applic.*, 31:49 – 60, 1996.

[85] R. Kornhuber. Adaptive monotone multigrid methods for some non–smooth optimization problems. Preprint No. 156, Weierstraß–Institut (WIAS), Berlin, 1995.

[86] R. Kornhuber and R. Roitzsch. Self adaptive computation of the breakdown voltage of planar pn–junctions with multistep field plates. In W. Fichtner and D. Aemmer, editors, *Simulation of Semiconductor Devices and Processes*, pages 535–543, Konstanz, 1991. Hartung–Gorre.

[87] R. Kornhuber and R. Roitzsch. Self adaptive finite element simulation of bipolar, strongly reverse biased pn-junctions. *Comm. Numer. Meth. Engrg.*, 9:243–250, 1993.

[88] R. Kornhuber and H. Yserentant. Multilevel methods for elliptic problems on domains not resolved by the coarse grid. In Keyes and Xu [80], pages 49–60.

[89] E. Laitinen and P. Neitaanmäki. On numerical simulation of the continuous casting process. *J. Eng. Math.*, 22:335–354, 1988.

[90] P. Leinen. *Ein schneller adaptiver Löser für elliptische Randwertprobleme*. PhD thesis, Universität Dortmund, 1990.

[91] E. Magenes, R.H. Nochetto and C. Verdi. Energy error estimates for a linear scheme to approximate nonlinear parabolic problems. *Math. Mod. Numer. Anal.*, 21:655–678, 1987.

[92] J. Mandel. Etude algébrique d'une méthode multigrille pour quelques problèmes de frontière libre. *C.R. Acad. Sci.*, 298:469–472, 1984. Ser. I.

[93] J. Mandel. A multilevel iterative method for symmetric, positive definite linear complementarity problems. *Appl. Math. Optimization*, 11:77–95, 1984.

[94] M. Markowich, C. Ringhofer and C Schmeiser. *Semiconductor Equations*. Springer, New York, 1990.

[95] S.F. McCormick. *Multilevel Projection Methods for Partial Differential Equations*. SIAM, Philadalphia, 1992.

[96] A. M. Meirmanov. *The Stefan Problem*. Walter de Gruyter, Berlin, 1992.

[97] P. Neitaanmäki and D. Tiba. *Optimal Control of Nonlinear Parabolic Systems*. Marcel Dekker, New York, 1994.

[98] R. H. Nocchetto, M. Paolini and C. Verdi. An adaptive finite element method for two–phase Stefan problems in two space dimensions. Part I Stability and error estimates. *Math. Comp.*, 57(195):73–108, 1991.

[99] R. H. Nocchetto, M. Paolini and C. Verdi. An adaptive finite element method for two–phase Stefan problems in two space dimensions. Part II Implementation and numerical experiments. *SIAM J. Sci. Stat. Comput.*, 12(5):1207–1244, 1991.

[100] R.H. Nochetto and C. Verdi. An efficient linear scheme to approximate parabolic free boundary problems: Error estimates and implementation. *Mathematics of Computation*, 51:27–53, 1988.

[101] O.A Oleinik. A method of solution of the general Stefan problem. *Sov. Math. Dokl.*, 1:1350–1354, 1960.

[102] O.A. Oleinik, A.S. Kalashnikov and ChZou Yui-Lin'. The Cauchy problem and boundary problems for equations of the type of nonstationary filtration. *Izv. Akad. Nauk SSSR Ser. Mat.*, 22:667–704, 1958.

[103] P. Oswald. Stable subspace splittings for Sobolev spaces and their applications. Forschungsergebnisse der Friedrich–Schiller–Universität Jena Math/93/7, Jena, 1993.

[104] Oswald P. *Multilevel Finite Element Approximation*. Skripten zur Numerik. Teubner, Stuttgart, 1994.

[105] J.F. Rodrigues. *Obstacle Problems in Mathematical Physics*. Number 134 in Mathematical Studies. North–Holland, Amsterdam, 1987.

[106] R. Roitzsch and R. Kornhuber. BOXES – A programm to generate triangulations from a rectangular domain description. Technical Report TR90-9, Konrad–Zuse–Zentrum (ZIB), Berlin, 1990.

[107] W.V. van Roosbroeck. Theory of flow of electrons and holes in germanium and other semiconductors. *Bell. Syst. Tech. J.*, 29:560–607, 1950.

[108] E. Rothe. Zweidimensionale parabolische Randwertaufgaben als Grenzfall eindimensionaler Randwertaufgaben. *Math. Ann.*, 102:650–670, 1930.

[109] L.I. Rubinstein. *The Stefan Problem*. AMS, Providence, 1971.

[110] J. Rulla and N. Walkington. Optimal rates of convergence for degenerate parabolic problems in two dimensions. *Preprint*, 1995.

[111] S. Selberherr. *Analysis and Simulation of Semiconductor Devices*. Springer, Wien, 1984.

[112] V.V. Shaidurov. Some estimates of the rate of convergence for the cascadic conjugate-gradient method. Preprint Nr. 4, Otto–von–Guericke–Universität, Magdeburg, 1994.

[113] G. Stampacchia. Formes bilinéaires coercitives sur les ensembles convexes. *C. R. Acad. Sci. Paris*, 258:4413–4416, 1964.

[114] J. Stefan. Ueber die Theorie der Eisbildung, insbesondere über Eisbildung im Polarmeere. *Ann. Phys. u. Chem*, 42:269–286, 1891.

[115] X.C. Tai. Domain decomposition for linear and nonlinear elliptic problems via function or space decomposition. In Keyes and Xu [80], pages 355–360.

[116] J.L. Vazquez. An introduction to the mathematical theorie of the porous medium equation. In M.C. Delfour, editor, *Shape Optimization and Free Boundaries*, volume 380 of *Mathematical and Physical Sciences*, pages 347–389, Dordrecht, 1992. Kluwer.

[117] Verein Deutscher Eisenhüttenleute. *Anhaltszahlen für die Wärmewirtschaft in Eisenhüttenwerken*, Düsseldorf, 1968. Verlag Stahleisen mbH.

[118] R. Verfürth. A posteriori error estimation and adaptive mesh–refinement techniques. *J. Comp. Appl. Math.*, 50:67–83, 1994.

[119] R. Verfürth. *A Review of a Posteriori Error Estimation and Adaptive Mesh–Refinement Techniques*. Skripten zur Numerik. Teubner, Stuttgart, to appear.

[120] R.L. Voller. Monoton einschließende Mehrgitterverfahren. *ZAMM*, 71, 1991.

[121] B. Wohlmut. *Adaptive Multilevel–Finite–Elemente Methoden zur Lösung elliptischer Randwertprobleme*. PhD thesis, TU München, 1995.

[122] J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Review*, 34:581–613, 1992.

[123] H. Yserentant. On the multi-level splitting of finite element spaces. *Numer. Math.*, 49:379–412, 1986.

[124] H. Yserentant. Two preconditioners based on the multilevel splitting of finite element spaces. *Numer. Math.*, 58:163–184, 1990.

[125] H. Yserentant. Hierarchical bases. In R.E. O'Malley, editor, *ICIAM 91*, pages 256–276, Philadelphia, 1992. SIAM.

[126] H. Yserentant. Old and new convergence proofs for multigrid methods. *Acta Numerica*, pages 285–326, 1993.

[127] X. Zhang. Multilevel Schwarz methods. *Numer. Math.*, 63:521–539, 1992.

[128] O.C. Zienkiewicz, J.P. de S.R. Gago and D.W. Kelly. The hierarchical concept in finite element analysis. *Computers & Structures*, 16:53–65, 1983.

[129] J. Zou. A new fast solver – Monotone MG Method (MMG). *J. Comp. Math*, 5:325–335, 1987.

Notation

$c, C, c_1, C_1, \ldots$	generic positive constants, usually not the same at different occurences
$\mathbb{R}$	real numbers
$x = (x_1, x_2)$	elements of $\mathbb{R}^2$
Ω, $\partial\Omega$	polygonal domain $\Omega \in \mathbb{R}^2$ with boundary $\partial\Omega$
$L^2(\Omega)$	Hilbert space of measurable, square integrable functions on the domain Ω
$\|\cdot\|_{L^2(\Omega)}$	norm on $L^2(\Omega)$: $\|v\|_{L^2(\Omega)} = \int_\Omega v^2 \, dx$
∂_i, ∇	generalized partial derivative ∂_i with respect to x_i and gradient $\nabla v = (\partial_1 v, \partial_2 v)$
$H^1(\Omega)$	Hilbert space of functions in $v \in L^2(\Omega)$ with first order derivatives $\partial_i v \in L^2(\Omega)$, $i = 1, 2$.
$\|\cdot\|_{H^1(\Omega)}$	norm in $H^1(\Omega)$: $\|v\|_{H^1(\Omega)} = (\|v\|^2_{L^2(\Omega)} + \sum_{i=1,2} \|\partial_i v\|^2_{L^2(\Omega)})^{1/2}$
$v_k \to v$, $k \to \infty$	convergence in H if not otherwise stated
$C_0^\infty(\Omega)$	space of infinitely differentiable functions with compact support in Ω
$H_0^1(\Omega)$	closure of $C_0^\infty(\Omega)$ in $H^1(\Omega)$
u, H	solution u and solution space H of the minimization problem (1.2.33), p. 22
$a(\cdot,\cdot)$, $\|\cdot\|$	elliptic bilinear form $a(v, v)$ and energy norm $\|v\|$, p. 22
$\mathcal{J}$	quadratic energy functional, p. 22
Φ	scalar convex function with properties (V1), (V2) and (V3) or (V3)', p. 23
ϕ	non–smooth convex functional generated by Φ, p. 22

$\partial\phi$, $\partial\Phi$	subdifferentials, p. 32
$\mathcal{T}_j$, $\mathcal{N}_j$, $\mathcal{E}_j$	triangulation $\mathcal{T}_j$ with interior nodes $\mathcal{N}_j$ and edges $\mathcal{E}_j$, p. 37
$\mathcal{S}_j$	space of piecewise linear finite elements with respect to the triangulation $\mathcal{T}_j$, p. 37
ϕ_j	approximation of ϕ with respect to $\mathcal{S}_j$, p. 37
Λ_j, $\lambda_p^{(j)}$	nodal basis Λ_j of $\mathcal{S}_j$ with elements $\lambda_p^{(j)}$, p. 37
u_j	solution of the discrete minimization problem (1.3.57), p. 37
$\bar{u}_j$	smoothed iterate, p. 52
$(M^\nu)_{\nu\geq 0}$, μ_l^ν	sequence of search directions $(M^\nu)_{\nu\geq 0}$ with elements μ_l^ν, see conditions (M1), p. 51, and (M2), (M3), p. 60
$\partial\Phi_l$	subdifferential of ϕ_j in direction of μ_l^ν, p. 50, 55
$\partial\Psi_l$	monotone approximation of $\partial\Phi_l$, p. 56
$\mathcal{N}_j^i(v)$, $\mathcal{N}_j^\bullet(v)$	discrete phase $\mathcal{N}_j^i(v)$ and critical nodes $\mathcal{N}_j^\bullet(v)$ of $v \in \mathcal{S}_j$, p. 59
$\mathcal{S}_j^\circ$	reduced subspace of $\mathcal{S}_j$, p. 63
$\Lambda_\mathcal{S}$, Λ	multilevel nodal basis $\Lambda_\mathcal{S}$ in $\mathcal{S}_j$ and corresponding search directions $\Lambda = (\Lambda_j, \Lambda_\mathcal{S})$, p. 69
$b_{\bar{u}_j}(w,\cdot) - f_{\bar{u}_j}$	local linearization of ϕ_j at $\bar{u}_j$, p. 62, 70
$\underline{\varphi}_j$, $\overline{\varphi}_j$	bounds for the neighborhood of $\bar{u}_j$ in which the local linearization is valid, p. 70
$\underline{\psi}_l$, $\overline{\psi_l}$	lower and upper bounds for the local corrections, p. 72
R_{k+1}^k	monotone restrictions, p. 74
$\tilde{\Lambda}_\mathcal{S}$, $\tilde{\lambda}_p^{(j)}$	truncated nodal basis $\tilde{\Lambda}_\mathcal{S}$ of $\mathcal{S}_j$ with elements $\tilde{\lambda}_p^{(j)}$, p. 79
$\mathcal{Q}_j$	space of piecewise quadratic finite elements with respect to the triangulation $\mathcal{T}_j$, p. 96
$\Lambda_\mathcal{Q}$, $\lambda_p^\mathcal{Q}$	nodal basis $\Lambda_\mathcal{Q}$ of $\mathcal{Q}_j$ with elements $\lambda_p^\mathcal{Q}$, p. 96
$e_\mathcal{Q}$	solution of the discrete defect problem (4.1.2), p. 96

Index