Technische Universität München
Lehrstuhl für mathematische Optimierung

A Class of Trust-Region Multilevel Methods

Boris Tobias von Loesch

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. Tim N. Hoffmann
Prüfer der Dissertation:	1.	Univ.-Prof. Dr. Michael Ulbrich
	2.	Univ.-Prof. Dr. Alfio Borzi Julius-Maximilians-Universität Würzburg

Die Dissertation wurde am 11.06.2012 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 05.12.2012 angenommen.

Abstract

In this thesis a class of trust-region multilevel methods for the solution of high-dimensional nonlinear optimization problems with convex constraints is investigated. Typical applications are discretizations of infinite-dimensional problems. Besides the actual objective function, the methods use models that can be evaluated more cheaply, for instance discretizations with less degrees of freedom. A comprehensive global convergence result is shown, where particular attention is paid to make all assertions largely independent of the problem's dimension. In a typical Sobolev space setting, it is further discussed under which conditions smoothing steps, that can be calculated cheaply, produce a sufficient descent. If these conditions are not met, the coarser models can be used instead. The application to typical problem classes is shown and numerical results of different examples, amongst others a 3D contact problem with nonlinear material model, confirm the excellent properties of the algorithm.

Zusammenfassung

Die vorliegende Arbeit befasst sich mit einer Klasse von Trust-Region Multilevelverfahren zum Lösen hochdimensionaler nichtlinearer Optimierungsprobleme mit konvexen Nebenbedingungen. Typische Anwendungsbeispiele sind Diskretisierungen unendlich-dimensionaler Optimierungsprobleme. Die untersuchten Verfahren verwenden neben der eigentlichen Zielfunktion auch günstiger zu berechnende Modelle, etwa Diskretisierungen mit weniger Freiheitsgraden. Für diese Klasse wird ein umfassendes globales Konvergenzresultat gezeigt. Hierbei wird besonders darauf geachtet, alle Aussagen weitgehend unabhängig von der Dimension der Probleme zu halten. Im weiteren Verlauf wird in dem typischen Fall, dass der zugrundeliegenden Raum ein Sobolev-Raum ist, untersucht, unter welchen Voraussetzungen numerisch günstige Glättungsschritte einen hinreichenden Abstieg liefern oder ob stattdessen Schritte auf einem anderen Modell gemacht werden sollten. Die Anwendung auf typische Problemklassen wird diskutiert und numerische Ergebnisse verschiedener Beispiele, unter anderem von einem 3D-Kontaktproblem mit nichtlinearen Materialmodell, bestätigen die guten Eigenschaften des Verfahrens.

Contents

1. Introduction

In this thesis we analyse a class of trust-region algorithms for the solution of convexly constrained optimization problems. Our main interest are objective functions that are discretized versions of a nonlinear functional which acts on an infinite dimensional space. A typical example of such an infinite dimensional problem from the calculus of variations is

$$\min_{u \in \mathcal{C}} \int_{\Omega} j(x, u(x), \nabla u(x)) \, \mathrm{d}x \tag{VP}$$

where $\mathcal{C}$ is a closed and convex subset of the Sobolev space $H^1(\Omega)$, $\Omega \subset \mathbb{R}^d$, and $j \colon \Omega \times \mathbb{R} \times \mathbb{R}^d$ a nonlinear function. These problems are typically large scale and therefore not well suited for standard optimization algorithms. Furthermore, the condition of the Hessians of these problems becomes large when the degrees of freedom grow due to a finer discretization. In this case, a simple steepest descent method requires more and more steps to reach a predefined precision. A simple example that illustrates this effect in one dimension was given in [Neu97, Chapter 2]. Contrary to that, it can be shown that (inexact) Newton's methods often behaves independent of the discretization [All86, WSD05]; but if no special care is taken, the effort for the computation of one Newton iteration grows more than proportional with the degrees of freedom. One of our major goals is to create an algorithm that does not exhibit this behaviour.

In the unconstrained case, the first-order optimality systems of problems of the type (VP) often corresponds to a (nonlinear) partial differential equation (PDE). For linear elliptic PDEs, *Multigrid* or *Multilevel algorithms* are computational optimal in the sense that the number of operations needed to reach a predefined precision depends only linearly on the degrees of freedom. These methods were first introduced in the early sixties by Fedorenko [Fed61, Fed64]. In the West, the first works on multilevel algorithms came from Brandt in 1973. First theoretical insights were given by the works of Nicolaides and Hackbusch. Since then multigrid methods attract a lot of attention and are still an active field of research. An elaborate description of the historical development till 1994 can be found in [Bra95].

Multigrid methods employ a hierarchy of discretizations with increasing degrees of freedom. The main observation that leads to the development of multigrid methods for linear systems was that cheap iterative solvers often effectively reduce the high frequencies of the error quickly but fail to diminish the low frequencies. The idea is to transfer the problem to a coarser grid where the error again has high frequency error components – in relation to the coarser discretization – which can be reduced effectively by the iterative methods. This is done in a recursively fashion on the complete hierarchy and leads to very effective solvers. A good overview over the theory and practice of multigrid methods can be found in the monographs [Hac85, Wes92, TOS01, BZ00].

Multigrid methods were also applied to solve nonlinear PDEs. Several different approaches are available to achieve this. The first one, often called *Newton Multigrid algorithm*, uses an outer

Newton iteration where the linear system is solved either directly by multigrid methods or with a preconditioned conjugated gradient algorithm where multigrid iterations are used as preconditioner. There is a large set of literature about these methods, for example [BR82, BVW03, Hac85]. These methods were also used to solve optimization problems where the multigrid algorithm is applied to the KKT-system [DW97, Kor01, DMS00] and, in particular, to solve PDE constrained optimization problems, see [BS09] and the references therein. Similarly, in [BHT09] the authors use multigrid methods to approximately solve the perturbed KKT-systems that occur in every iteration of an interior point algorithm.

Another approach used for unconstrained and constrained convex optimization problems are subspace correction methods where the function is successively minimized over a large series of simple – often one dimensional – subspaces [Tai03, Bad06, BTW03, Kor94, GK09b]. The basis for these algorithms is a different interpretation of multilevel algorithms, namely as *subspace correction algorithms* [Xu92, Yse93]. In [Tai03, Bad06], the authors show that the number of iterations needed to reach a given precision is bounded independently of the fineness of the discretization if the problem is unconstrained and bounded by a constant that depends weakly on the degrees of freedom if the problem is constrained by simple pointwise bounds.

Finally, there are methods where the multigrid iteration is directly applied to the nonlinear problems. The two main methods are FAS (Full Approximation Scheme), proposed in [Bra77], and NMGM (Nonlinear Multigrid Method) [Hac85, HR89]. Extensions were also used to solve variational inequalities [Hop87, BC83, Man84]. Based on these methods, in [Nas00] the MG/Opt algorithm was introduced. Here, for the first time, we have a truly nonlinear multigrid method for unconstrained convex optimization problems that is independent of a PDE setting. MG/Opt formulates the FAS method for optimization problems and uses a line search algorithm to ensure global convergence. An improved variant of the algorithm was introduced in [LN05] where the length of the lower-level steps is bounded similar to a trust-region approach. In the short note [Bor05], the global convergence of MG/Opt was shown for strictly convex problems by applying the theory of [HR89]. In [WG09] a more elaborate convergence theory of an algorithm similar to MG/Opt was made, which also includes an estimate of the total number of iterations needed to obtain a given precision if the objective function is uniformly convex. However, this estimate depends on the condition of the Hessian matrices and is of the same order as the number of steps needed in a steepest descent algorithm.

Instead of a line search, the algorithm RMTR (Recursive Multiscale Trust-Region algorithm) [GST08] uses a trust region for globalization. A comprehensive convergence analysis was made, which does not need the convexity of the objective function. For the first time, the usage of coarser models was restricted to cases where one can expect good steps. We emphasize this point since it will become important in our analysis. The algorithm was extended to box-constrained problems in [GMTWM08, GMS$^+$10]. A variant which needs less differentiability assumptions was considered later in [GK09a].

These more recent multilevel optimization algorithms were all formulated in a typical Euclidean, finite dimensional setting, which allows a broad usage for many optimization problems as long as a suitable level hierarchy is available. The often infinite dimensional structure of the underlying problem is not taken into account. Therefore, the constants that appear in the statements often depend on the discretization, which leads to estimates that are highly level-dependent. This distinguishes the analysis of these class of optimization algorithms from other multilevel

methods where the special structure is heavily used to show level-independent convergence behaviour.

In this thesis we try to bridge this gap and bring – at least partially – these different approaches closer together. The main algorithm we analyze in this thesis is based on RMTR and RMTR$_\infty$, which is the version for box-constrained problems. However, we generalize these algorithms in various ways. Not only classical multilevel settings fit in our framework but also domain decomposition methods or a combination of both are possible. In theory, we are not limited to bound constrained problems, but allow general convex feasible sets. But most importantly, we analyse the algorithm with the infinite dimensional setting in mind. That means we work directly in abstract Hilbert spaces whenever possible and carefully pay attention whether constants depend on the dimension of the problem. Furthermore, we later restrict ourselves to a more concrete setting and consider cheap-to-calculate multigrid smoothers for the calculation of steps and analyse the descent that we obtain. To this end, we will use results from the theory of subspace correction methods.

We end this introduction by summarizing the contents of the upcoming chapters.

We start Chapter 2 by introducing the abstract setting we are going to consider and present a trust-region multilevel algorithm for convexly constrained problems. We continue to show the convergence to first-order critical points under quite general assumptions on the function and the hierarchy. The theory is a lot more general than it is needed for the chapters that follows where we restrict ourselves to a more concrete setting. However, we hope to identify the important assumptions that must be satisfied to show global convergence and that allows one to easily use the theory for other problem classes.

In Chapter 3 we consider unconstrained problems in a setting that one typically has if the function hierarchy was created by discretization of an infinite dimensional problem with finite elements. This variational setting will be introduced first, and an important connection between smoothness and estimates of certain norms will be pointed out. Then we analyze the descent which we obtain by typical smoothing methods for convex and non-convex problems. For this we use the abstract theory of subspace corrections algorithms. We finish this chapter with some remarks about the concrete implementation of the smoothers.

Chapter 4 considers problems where the feasible set is a closed and convex proper subset of the whole space. We start by introducing a continuous stationarity measure that can be calculated with a reasonable effort. Similarly to the second chapter, we then analyse the descent produced of various smoothing methods. Finally, we turn to a special class of feasible sets, boxes, and show how the abstract choices of lower-level sets introduced in Chapter 2 can be implemented. In this case we will also present an active-set strategy which can greatly improve the convergence speed of the method.

In Chapter 5 we will consider concrete classes of infinite dimensional problems and establish the various assumptions that we made so far.

Finally, in the last chapter, we will show convincing numerical results of the algorithm on selected 2D and 3D examples and different choices of parameters.

2. A multilevel trust-region algorithm

In this chapter we will introduce a multilevel trust-region algorithm which is applicable to a wide range of problems. It is evolved from the algorithms RMTR [GST08] and RMTR$_\infty$ [GMTWM08]. Before we state the algorithm, we will give a motivation of its ingredients and show how to choose them in some common settings. We finally show the global convergence to first-order stationary points.

We start by introducing some basic notation that we will use subsequently.

2.1. Notation

Let X be a normed vector space over $\mathbb{R}$. The *dual space*, X^* denotes the space of all bounded and linear mappings of X to $\mathbb{R}$, $\mathcal{L}(X, \mathbb{R})$. Instead of the notation $g(x)$ for $g \in X^*$ we often use the *dual pairing*

$$\langle g, x\rangle_{X^*,X}.$$

In most cases, we omit the spaces in the above notation and just write $\langle g, x\rangle$. X^* equipped with the norm

$$\|g\|_{X^*} := \sup_{\substack{x\in X\\ \|x\|_X=1}} \langle g, x\rangle_{X^*,X}$$

is a Banach space. It follows from the definition of the dual norm that

$$\langle g, x\rangle_{X^*,X} \leq \|x\|_X \|g\|_{X^*}. \tag{2.1}$$

By $\mathcal{L}(X, Y)$ we denote the space of linear continuous operators between two normed vector spaces X and Y. Every operator $P \in \mathcal{L}(X, Y)$ is bounded, i.e., there exists a positive constant M such that $\|Px\|_Y \leq M\|x\|_X$ for all $x \in X$. The operator norm on this space is given by

$$\|P\|_{X,Y} := \sup_{\|x\|_X=1} \|Px\|_Y.$$

The *dual* or *adjoint* operator of $P \in \mathcal{L}(X, Y)$ is denoted by $P^* \in \mathcal{L}(Y^*, X^*)$ and satisfies

$$\langle g, Px\rangle_{Y^*,Y} = \langle P^*g, x\rangle_{X^*,X} \quad \text{for all } x \in X,\ g \in Y^*.$$

If X is reflexive, $Y = X^*$ and P fulfills

$$\langle Py, x\rangle_{X^*,X} = \langle Px, y\rangle_{X^*,X} \quad \text{for all } x, y \in X,$$

we call the operator P *symmetric* or *self-adjoint*.

2.1.1. Lebesgue and Sobolev spaces

Let $\Omega \subset \mathbb{R}^d$, $d \geq 1$, be a domain with Lipschitz-continuous boundary $\partial\Omega$. We use the standard notations $L^p(\Omega)$ with $1 \leq p < \infty$ for the Lebesgue spaces consisting of p-th power integrable functions and $L^\infty(\Omega)$ for the space of essentially bounded functions. Let $W^{m,p}(\Omega) \subset L^p(\Omega)$ be the set of all functions having weak derivatives $D^\alpha u \in L^p(\Omega)$ for $|\alpha| \leq m$:

$$W^{m,p}(\Omega) := \{u \in L^p(\Omega) \,|\, D^\alpha u \in L^p(\Omega) \text{ for } |\alpha| \leq m\}.$$

The set $W^{m,p}(\Omega)$ with the norm

$$\begin{aligned} \|u\|_{W^{m,p}(\Omega)} &:= \left(\sum_{|\alpha| \leq m} \|D^\alpha u\|^p_{L^p(\Omega)} \right)^{1/p}, \quad p \in [1, \infty), \\ \|u\|_{W^{m,\infty}(\Omega)} &:= \sum_{|\alpha| \leq m} \|D^\alpha u\|_{L^\infty(\Omega)}. \end{aligned}$$

forms a Banach space and is called *Sobolev space of index* (m, p). In the special case $p = 2$, $H^m(\Omega) := W^{m,2}(\Omega)$ with the inner product

$$(u, v)_{H^m(\Omega)} := \sum_{|\alpha| \leq m} (D^\alpha u, D^\alpha v)$$

is a Hilbert space. We will often work in the space $H^1_0(\Omega)$, which can be characterized by

$$H^1_0(\Omega) = \{u \in H^1(\Omega) |\ \operatorname{tr} u = 0 \text{ on } \partial\Omega\}.$$

Here, $\operatorname{tr}\colon H^1(\Omega) \to L^2(\partial\Omega)$ is a continuous linear mapping with $\operatorname{tr} v = v|_{\partial\Omega}$ for all $v \in C^1(\overline{\Omega})$ called the trace operator. This mapping exists under the assumptions on the domain Ω. When no confusion arises, we simply write $\operatorname{tr} u = u$.

2.1.2. Gelfand triple

Let $\mathcal{V}$ be a reflexive Banach space that densely and continuously embeds into a Hilbert space $\mathcal{U}$. By the Riesz representation theorem we can identify $\mathcal{U}$ with $\mathcal{U}^*$ by means of the embedding $\iota_\mathcal{U}\colon \mathcal{U} \to \mathcal{U}^*$, $u \mapsto (\cdot, u)_\mathcal{U}$. Then $\mathcal{U}^* = \mathcal{U}$ is embedded continuously and densely into the dual space $\mathcal{V}^*$. The chain $\mathcal{V} \hookrightarrow \mathcal{U} \hookrightarrow \mathcal{V}^*$ is called a *Gelfand triple*. The continuous extension of the scalar product $(\cdot, \cdot)_\mathcal{U}$ to $\mathcal{V} \times \mathcal{V}^*$ results in the dual form $\langle \cdot, \cdot \rangle_{\mathcal{V}^*, \mathcal{V}}$. Hence, we use the notation $(g, v)_\mathcal{U}$ for $v, g \in \mathcal{U}$ as well as $g \in \mathcal{V}^*$, $v \in \mathcal{V}$.

An example for a Gelfand triple is

$$H^1_0(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow H^{-1}(\Omega) := H^1_0(\Omega)^*.$$

2.2. Problem setting

In this chapter we present a multilevel trust-region algorithm for convexly constrained problems and prove global convergence. Let C_r be a closed and convex subset of a Banach space $\mathcal{V}_r$. We consider the problem

$$\min_{x_r \in C_r} f_r(x_r). \tag{2.2}$$

We assume that the function $f_r \colon \mathcal{V}_r \to \mathbb{R}$ is continuously differentiable and that the second-order Gateï¿½ux derivative exists and the mappings $x_r \mapsto f_r''(x_r)[h,h]$ are continuous for all $h \in \mathcal{V}_r$. This is satisfied if, for instance, f_r is twice continuously differentiable. Since $f_r''(x_r) \in \mathcal{L}(\mathcal{V}_r, \mathcal{V}_r^*)$, we also use the notation $\langle f_r''(x_r)h, h\rangle$.

We are interested in cases where (2.2) is a large-scale optimization problem and where besides the objective function f_r there are auxiliary functions

$$f_i \colon \mathcal{W}_i \times \mathcal{V}_i \to \mathbb{R}, \quad i = 1, \dots, r-1$$

defined on – normally lower dimensional – spaces $\mathcal{V}_i$ and $\mathcal{W}_i$, which are somehow connected to f_r. For every $x_i \in \mathcal{W}_i$, the functions $f_i(x_i, \cdot) \colon \mathcal{V}_i \to \mathbb{R}$ are assumed to have the same differentiability properties as f_r. Each of these functions serves as a model of f_r at a point x_r and we suppose that evaluating the auxiliary functions is cheaper in terms of computational effort than evaluating f_r. Each time a lower-level function f_i is used to obtain a new iterate, the point $x_i \in \mathcal{W}_i$ is fixed. This allows us to use different spaces $\mathcal{W}_i$ and $\mathcal{V}_i$ for the "development points" and the search directions. However, in many applications the spaces $\mathcal{W}_i$ are equal to $\mathcal{V}_i$ and $f_i(x_i, v_i) := \hat{f}_i(x_i + v_i)$ with $\hat{f}_i \colon \mathcal{V}_i \to \mathbb{R}$ holds.

A typical example is when the spaces $\mathcal{V}_i$ form a nested sequence of finite dimensional spaces with increasing dimension, e.g., constructed by a successive refinement process and the functions f_i are approximations of f_r on $\mathcal{V}_i$. This is similar to a classical multigrid setting. Throughout this work we are mostly concerned with such multigrid hierarchies and hence we will often use the terms *coarse* and *fine* to distinguish between the spaces $\mathcal{V}_i$.

Besides multigrid hierarchies, other choices of $\mathcal{V}_i$ and f_i are possible. Domain decomposition methods like the alternating Schwarz method use a divide and conquer methodology to solve problems (typically PDE's) that are defined on a large domain by splitting it into smaller parts. On each subdomain an approximation of the original problem is solved and its solutions are merged to obtain an approximate solution of the problem on the whole domain. Assume that $\mathcal{V}_r$ is a finite dimensional function space over a domain Ω. We split Ω into (not necessarily disjunct) subdomains $\{\Omega_i\}_i$ and define $\mathcal{V}_i$ as a suitable function space over Ω_i for all i. Since the elements of $\mathcal{V}_i$ are only defined on a part of the whole domain, the functions f_i must be chosen suitably to approximate f_r on $\mathcal{V}_i$. A concrete choice is given in Example 2.1.

Combinations of multilevel and domain decomposition approaches are also possible, e.g., an overlapping domain decomposition approach with an additional coarse space.

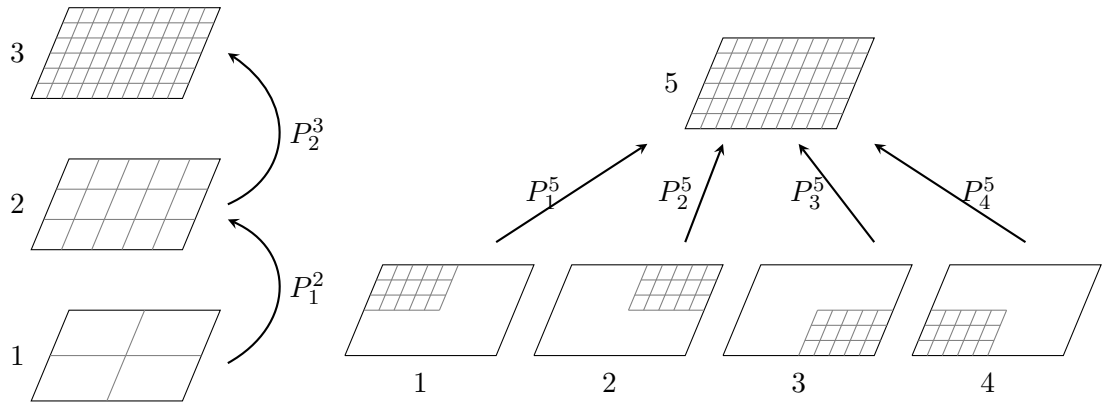

Figure 2.1.: Two examples of a hierarchy

2.2.1. Function hierarchies

To use the auxiliary functions, they must be connected to f_r and proper transfer operators between the spaces must exists.

In order to treat all levels the same, we set $\mathcal{W}_r = \mathcal{V}_r$ and define the function $f_r \colon \mathcal{W}_r \times \mathcal{V}_r \to \mathbb{R}$ by $f_r(x_r, v_r) := f_r(x_r + v_r)$, where the right-hand side is given by the objective function of problem (2.2). Although both functions have the same name, no confusion should arise since the number of arguments is different.

To describe the hierarchy, we define for every level index i the set of direct children nodes $N(i) \subset \{1, \dots, r\}$, which can be empty. We use the relation $j \prec i$ to indicate that there is a path connecting level j and level i. More precisely, there is a sequence or chain of levels $(j, l_1, \dots, l_m, i)$ such that

$$j \in N(l_1), \quad l_{k-1} \in N(l_k),\ k = 2, \dots, m \quad \text{and } l_m \in N(i). \tag{2.3}$$

From the definition it is clear that $\prec$ is transitive, i.e., from $j \prec i$ and $i \prec l$ follows $j \prec l$.

We require the hierarchy to be a tree with level r as root node in the sense of graph theory. This means especially that the path between two levels j, i with $j \prec i$ is unique and that $i \prec r$ for all $i = 1, \dots, r-1$. Furthermore a tree is circle free, i.e., from $j \prec i$ follows $i \not\prec j$. This and the fact that the number of levels is finite imply that every chain between two levels is finite. By $\sharp i$ we denote the maximum length of a level-chain that ends at level i, i.e.,

$$\sharp i := \max\left\{0, \max_{s \in S(i)} |s|\right\}, \quad S(i) := \left\{(l_1, \dots, l_m) \,|\, l_m \in N(i),\ l_{k-1} \in N(l_k)\ \forall k = 2, \dots, m\right\}.$$

Here, $|s|$ denotes the number of elements in s. If $N(i) = \emptyset$, we get $\sharp i = 0$. For $j \in N(i)$ it is easy to see that $\sharp j \leq \sharp i - 1$ and hence $\sharp r \geq \sharp i$ for all levels i.

Remark 2.1 We will often assume a multigrid level structure, where the levels are numbered increasingly from the coarsest to the finest. In this case, we set $N(i) = \{i - 1\}$, $i = 2, \dots, r$, and $N(1) = \emptyset$. Then $j \prec i$ is equivalent to $j < i$ and $\sharp i = i - 1$.

To connect the levels, for every pair (i, j) with $j \in N(i)$ there must be a *restriction operator*

$$R_i^j \colon \mathcal{W}_i \times \mathcal{V}_i \to \mathcal{W}_j$$

and a linear and continuous *prolongation operator*

$$P_j^i \colon \mathcal{V}_j \to \mathcal{V}_i.$$

As a natural extension, we define a prolongation for every pair of levels (i, k) with $i \prec k$ by successive prolongation from i to k:

$$\mathcal{P}_i^k = P_{l_m}^k \cdots P_{l_1}^{l_2} P_i^{l_1}$$

where $(i, l_1, \ldots, l_m, k)$ describes the unique sequence of levels from i to k in the sense of (2.3).

In general, we allow that the prolongation operators P_j^i are not fixed, but depend on the current iterate on level i. An example for this will be the active-set strategy for bound constrained problems where we use slight modifications of the standard prolongation operators. Of course this will also affect the operators $\mathcal{P}_i^k$, which then depend on all iterates of the intermediate levels. To simplify the notation we omit an additional iteration index.

The following examples will show how to concretely choose the spaces, the auxiliary functions and the transfer operators in two different settings.

Example 2.1 (Obstacle problem) Let us consider a membrane with uniform tension τ attached to the boundary $\partial\Omega$ of a domain $\Omega \subset \mathbb{R}^2$ above a rigid obstacle $\varphi \in H^2(\Omega)$ with $\varphi \leq 0$ on $\partial\Omega$. A vertical force with density τf, $f \in L^2(\Omega)$, is acting on the membrane. If we consider only small strains, the vertical displacement u of the membrane is the function that minimizes the membrane energy

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

over the set $U := \{u \in H_0^1(\Omega) \,|\, u \geq \varphi \text{ a.e. in } \Omega\}$ of admissible displacements. Since U is closed and convex, it follows directly from the Lax-Milgram lemma (see for instance [Bra07]) that this problem has a unique solution.

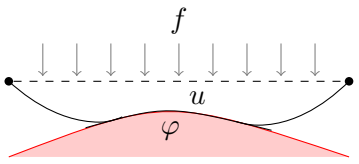

Figure 2.2.: Obstacle problem

In the following, let $\overline{\Omega}$ be polygonal. To discretize the problem, we start with a triangulation $\mathcal{T}_1$ of $\overline{\Omega}$ with simplices t of diameter less than h_1. Starting from this coarse triangulation, a sequence of triangulations $\mathcal{T}_2, \ldots, \mathcal{T}_r$ of $\overline{\Omega}$ is created by uniform refinement with mesh sizes $h_2, \ldots, h_r$. This

ensures that the sets of nodes $\mathcal{N}_i,\ i = 1, \dots, r$, which consist of all vertices of $\mathcal{T}_i$, are nested. On the triangulations we define conforming finite element spaces

$$S_i := \{u \in C^0(\overline{\Omega}) \,|\, u = 0 \text{ on } \partial\Omega,\ u \text{ restricted to } t \text{ is affine linear for all } t \in \mathcal{T}_i\}.$$

Since the sets of nodes $\mathcal{N}_i$ are nested, $S_1 \subset S_2 \subset \cdots \subset S_r \subset H_0^1(\Omega)$ holds. We now consider the discrete problem

$$\min_{u_r \in C_r} J_r(u_r) := \frac{1}{2} \int_\Omega \|\nabla u_r\|^2 \,\mathrm{d}x - \int_\Omega f u_r \,\mathrm{d}x$$

with the feasible set $C_r = U_r := \{u \in S_r \,|\, u \geq \varphi_r \text{ in } \Omega\}$ where $\varphi_r \in S_r$ is the nodal interpolant of the obstacle φ satisfying $\varphi_r(x) = \varphi(x)$ for $x \in \mathcal{N}_r$. Notice that in general $U_r \not\subset U$. In the same way, we can define functionals $J_i \colon S_i \to \mathbb{R}$ on the coarser grids.

If Ω is convex, one can show the estimate $\|u^* - u_r^*\|_{H^1(\Omega)} \leq C(u^*, f, \varphi) h_r$ for the error between the continuous solution u^* and the solution u_r^* of the discretized problem, where h_r is the maximum diameter of the triangles in $\mathcal{T}_r$ (cf. [Cia78, Section 5.1]). The constant $C(u^*, f, \varphi)$ is independent of the mesh size.

We now construct a valid multilevel hierarchy according to Section 2.2.1. The child sets $N(i)$, $i = 1, \dots, r$, are set as in Remark 2.1. We show two different ways how to define the spaces $\mathcal{V}_i$ and $\mathcal{W}_i$ and the transfer operators between adjacent levels:

1. Set $\mathcal{V}_r = S_r,\ f_r(v_r) := J_r(v_r)$ and

   $$\mathcal{V}_i = \mathcal{W}_i = S_i,\ f_i(x_i, v_i) := J_i(x_i + v_i) \text{ for } i = 1, \dots, r-1.$$

   Since the spaces $\mathcal{V}_i$ are nested, we can use the identity $\mathrm{id}_{i-1} \colon \mathcal{V}_{i-1} \to \mathcal{V}_i$ as prolongation operator P_{i-1}^i. An element $u_i \in S_i$ can be restricted to S_{i-1} by means of a nodal interpolation $I_{i-1} \colon H_0^1(\Omega) \to S_{i-1}$, i.e., $u_{i-1} = I_{i-1} u$ is the unique element that satisfies $u_{i-1}(x_{i-1}) = u(x_{i-1})$ for all $x_{i-1} \in \mathcal{N}_{i-1}$. The restriction operators are now defined by $R_i^{i-1}(x_i, v_i) := I_{i-1}(x_i + v_i)$.

2. Alternatively, one can use the coarser spaces S_i together with the functional on level r. Set $\mathcal{V}_r = S_r$, $f_r(v_r) := J_r(v_r)$ and

   $$\mathcal{V}_i = S_i,\ \mathcal{W}_i = S_r,\ f_i(x_i, v_i) := J_r(x_i + v_i) \text{ for } i = 1, \dots, r-1.$$

   In this case, we can use the identity by means of $R_i^{i-1}(x_i, v_i) := x_i + v_i$ as restriction and, as in the first case, the identity as prolongation operator.

The second approach has the disadvantage that in general the evaluation of f_i is as expensive as of f_r.

Alternatively, we can also build a hierarchy for an overlapping domain decomposition approach. For this let the domain Ω be partitioned into $r-1$ polygonal subdomains Ω_i such that $\Omega = \bigcup_{i=1}^{r-1} \Omega_i$ holds. The intersection of two neighbouring subdomains is assumed to be non-empty. We set $N(r) = \{1, \dots, r-1\}$ and $N(i) = \emptyset$ for $i = 1, \dots, r-1$. For simplicity, we assume that we have a triangulation $\mathcal{T}$ of $\overline{\Omega}$ that is consistent with the triangulations of the subdomains, i.e., there are

subsets $\mathcal{T}_i$ of $\mathcal{T}$ such that $\overline{\Omega}_i = \bigcup_{t\in\mathcal{T}_i} t$. The set of nodes of $\mathcal{T}_i$ is denoted by $\mathcal{N}_i$. On each Ω_i we define a finite dimensional function space by

$$S_i := \{u \in C^0(\overline{\Omega}_i) \,|\, u \text{ restricted to } t \text{ is affine linear for all } t \in \mathcal{T}_i,\ u = 0 \text{ on } \partial\Omega \cap \partial\Omega_i\}$$

for $i = 1, \dots, r-1$, and by $S_r \subset H_0^1(\Omega)$ the linear finite element space on Ω. Furthermore, we set $S_{i,0} := S_i \cap H_0^1(\Omega_i)$ for $i = 1, \dots, r-1$. There are natural extension operators $P_i \colon S_{i,0} \to S_r$, which take local functions on Ω_i with zero boundary conditions and extend them by zero on $\Omega \setminus \Omega_i$:

$$P_i \colon S_{i,0} \to S_r, \quad (P_i u_i)(x) := \begin{cases} u_i(x) & \text{if } x \in \Omega_i, \\ 0 & \text{if } x \in \Omega \setminus \Omega_i. \end{cases}$$

Similarly, we define for the restriction of elements $u_r \in S_r$ the operators $R_i \colon S_r \to S_i$, $R_i u_r = u_r|_{\Omega_i}$. Both P_i and R_i are linear and well-defined since the triangulations are consistent. In the case of nonmatching grids, the operators R_i and P_i can be defined by interpolation.

As in the multilevel scenario, at least two different possible constructions of hierarchies are possible:

1. Set $\mathcal{V}_r = S_r$, $\mathcal{W}_i = S_i$ and $\mathcal{V}_i = S_{i,0}$ for $i = 1, \dots, r-1$. Define f_i by

   $$f_i(w_i, v_i) := \frac{1}{2}\int_{\Omega_i} \|\nabla(w_i + v_i)\|^2 \,\mathrm{d}x - \int_{\Omega_i} f \cdot (w_i + v_i)\,\mathrm{d}x,$$

   the prolongations by $P_i^r = P_i$ and the restrictions by $R_r^i = R_i$.

2. Set $\mathcal{V}_r = S_r$, $\mathcal{W}_i = S_r$ and $\mathcal{V}_i = S_{i,0}$ for $i = 1, \dots, r-1$. Define the functions f_i by $f_i(x_i, v_i) := J_r(x_i + P_i v_i)$. As in the multilevel case, the identity can be used as restrictions R_r^i. The prolongations are chosen as in the first setting.

Example 2.2 (Obstacle Bratu problem) We consider the nonlinear problem suggested in [Mor90] given by

$$-\Delta u \le \lambda e^u \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega, \tag{2.4}$$
$$u \le \psi \text{ and } (-\Delta u - \lambda e^u)(u - \psi) = 0 \text{ in } \Omega \tag{2.5}$$

where $\Omega = (0,1)^2$ is the unit square, $\lambda \in \mathbb{R}$ a parameter and $C^0(\Omega) \ni \psi \ge 0$ an obstacle function. We introduce a regular grid with mesh width (h_x, h_y) on Ω and $\partial\Omega$ by

$$\Omega_h := \{(x,y) \in \Omega \,|\, x = i \cdot h_x,\ y = j \cdot h_y,\ i,j \in \mathbb{Z}\},$$
$$\partial\Omega_h := \{(x,y) \in \partial\Omega \,|\, x = i \cdot h_x,\ y = j \cdot h_y,\ i,j \in \mathbb{Z}\}.$$

For simplicity, we assume that $h = h_x = h_y$ and that for every $(x,y) \in \Omega_h$ the neighbouring points $(x \pm h_x, y \pm h_y)$ are contained in $\Omega_h \cup \partial\Omega_h$. We are interested in approximate solutions $u_h \colon \Omega_h \cup \partial\Omega_h \to \mathbb{R}$ of (2.4) on the grid Ω_h. We discretize the system by means of

$$-\Delta_h u_h \le \lambda e^{u_h}, \text{ in } \Omega_h, \quad u_h = 0 \text{ on } \partial\Omega_h,$$
$$u_h \le \psi_h \text{ and } (-\Delta_h u_h - \lambda e^{u_h})(u_h - \psi_h) = 0 \text{ in } \Omega_h.$$

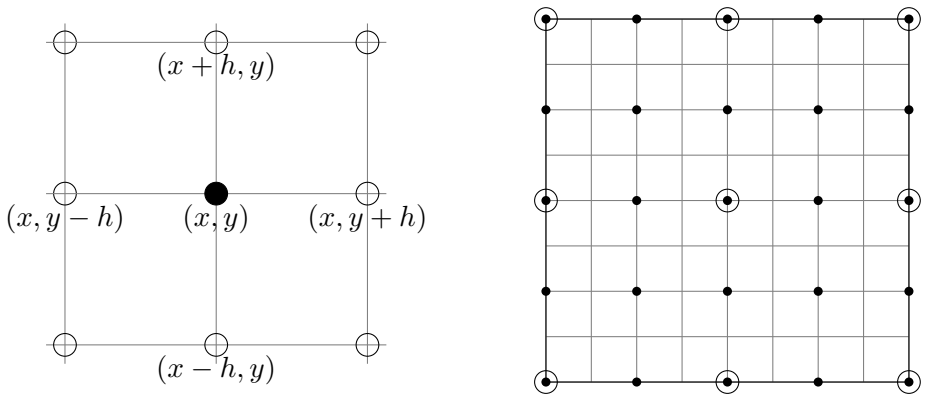

Figure 2.3.: Stencil notation and different grids for $h_3 = 1/8$, $h_2 = 1/4$ and $h_1 = 1/2$.

$\psi_h \colon \Omega_h \to \mathbb{R}$ is a grid function with $\psi_h(x,y) = \psi(x,y)$ for $(x,y) \in \Omega_h$. For the discretization Δ_h of the Laplace operator, we use the classical *five-point approximation*:

$$\begin{aligned} -(\Delta_h u_h)(x,y) &= \frac{1}{h^2}\left[4u_h(x,y) - u_h(x-h,y) - u_h(x+h,y) - u_h(x,y-h) - u_h(x,y+h)\right] \\ &= \frac{1}{h^2}\begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}_h u_h(x,y). \end{aligned}$$

The formula of the last line uses the descriptive *stencil notation* (cf., e.g., [TOS01, Wes92]). By ordering the values of Ω_h lexicographically, there is a unique representation of a grid function $u_h \in \Omega_h$ by a vector $\tilde{u}_h \in \mathbb{R}^n$ with length $n = |\Omega_h|$. Other orderings of the grid points, e.g., red-black ordering, are also possible. In the following, we will not distinguish between the grid function and its vector representation and simply write u_h for $\tilde{u}_h$ and ψ_h for $\tilde{\psi}_h$ when no confusion can arise.

The operator $-\Delta_h$ is linear and can be represented by a symmetric matrix $h^{-2}A_h \in \mathbb{R}^{n\times n}$. Here, the values of u_h on the boundary of Ω are considered to be zero, which is compatible with the Dirichlet boundary condition. Finally we arrive at the following nonlinear system in $\mathbb{R}^n$:

$$h^{-2}A_h u_h \le \lambda e^{u_h},\ u_h \le \psi_h \text{ and } \left(h^{-2}A_h u_h - \lambda e^{u_h}, u_h - \psi_h\right) = 0. \tag{2.6}$$

Here, $(\cdot,\cdot)$ denotes the standard inner product on $\mathbb{R}^n$. It is well known that A_h is irreducibly diagonal dominant and hence positive definite [Hac92, Criterion 4.3.24]. If $\lambda \le 0$, the nonlinear operator

$$\Phi_h(u_h) := h^{-2}A_h u_h - e^{\lambda u_h}$$

is *monotone* in the sense that

$$(\Phi_h(u_h) - \Phi_h(\bar{u}_h), u_h - \bar{u}_h) \ge 0 \quad \forall\, u_h, \bar{u}_h \in \mathbb{R}^n.$$

Φ_h is the gradient of the function $\phi_h(u_h) := \frac{1}{2}h^{-2}\,(u_h, A_h u_h) - \lambda \sum_{i=1}^n e^{u_h^i}$ and since Φ_h is monotone, ϕ_h is a convex function.

In order to show that solving the discretized problem is equivalent to an optimization problem, we use the following well-known characterization of the solutions of bound constrained problems:

Lemma 2.1 *Let $\mathcal{B} := [l, u] \subset \mathbb{R}^n$ be a box with bounds $l \in \mathbb{R}^n \cup \{-\infty\}$ and $u \in \mathbb{R}^n \cup \{\infty\}$ [1]. If x^* is a local solution of*

$$\min_{x \in \mathcal{B}} f(x)$$

and f is differentiable in x^*, then

$$x^* \in \mathcal{B} \quad \text{and} \quad \nabla f(x^*)^i \begin{cases} = 0 & \text{for } l^i < (x^*)^i < b^i, \\ \leq 0 & \text{for } u^i = (x^*)^i, \\ \geq 0 & \text{for } l^i = (x^*)^i \end{cases} \qquad \text{for } i = 1, \ldots, n \tag{2.7}$$

is satisfied. Moreover, if $l \equiv -\infty$ (analogously: $u \equiv \infty$), (2.7) can equivalently be written as

$$x^* \in \mathcal{B}, \quad \nabla f(x^*) \leq 0 \quad \text{and} \quad (u - x^*, \nabla f(x^*)) = 0. \tag{2.8}$$

If f is convex and (2.7) or (2.8) is satisfied for $x^* \in \mathcal{B}$, then x^* is a global solution of the minimization problem.

A proof can be found for instance in [UUH99, Thm. 4.1].

Using ϕ_h, (2.6) can be written as

$$\nabla \phi_h(u_h) \leq 0, \ u_h \leq \psi_h \text{ and } (\nabla \phi_h(u_h), u_h - \psi_h) = 0.$$

Accordingly, Lemma 2.1 yields that solving (2.6) is equivalent to finding a solution of the problem

$$\min_{u_h \in \mathbb{R}^n} \phi_h(u_h) \quad \text{subject to} \quad u_h \leq \psi_h,$$

which is a bound constrained optimization problem in $\mathbb{R}^n$.

Let us now assume that we have different grids Ω_{h_i}, $i = 1, \ldots, r$, and the grid-sizes satisfy the relation $2h_{i+1} = h_i$ for $i = 1, \ldots, r-1$; cf. Figure 2.3 for an example. We denote the associated coefficient spaces by $\mathbb{R}^{n_i}$. A grid function $u_{h_i} \in \Omega_{h_i}$ can be prolongated to $\Omega_{h_{i+1}}$ by standard bilinear interpolation (Figure 2.4). In stencil notation we can write this operator as

$$\left[P_i^{i+1}\right] = \frac{1}{4} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}_{h_i \to h_{i+1}}.$$

If we use a properly scaled inner product on $\mathbb{R}^{n_i}$,

$$(\cdot, \cdot)_{h_i} := h_i^2 (\cdot, \cdot),$$

the adjoint operator relative to $(\cdot, \cdot)_{h_i}$ and $(\cdot, \cdot)_{h_{i+1}}$ is the *full weighting* operator, which in stencil notation reads

$$\left[(P_{i+1}^i)^*\right] = \frac{1}{16} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}_{h_{i+1} \to h_i}.$$

[1] The notation $[l, u] = \{x \in \mathbb{R}^n \,|\, l \leq x \leq u\}$ is meant componentwise.

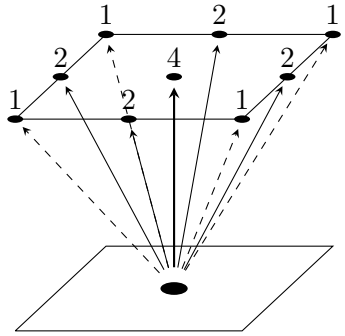

Figure 2.4.: Prolongation by bilinear interpolation

To restrict a point we can use the full weighting operator or a simple injection which is defined by

$$I_{i+1}^{i}(u_{h_{i+1}})(x,y) = u_{h_{i+1}}(x,y) \quad \forall\,(x,y) \in \Omega_{h_i}.$$

In our trust-region algorithm we compare the reduction achieved by a step on a lower grid with the reduction of the prolongated step. This suggests that the functions of our hierarchy should be scaled such that $\phi_{h_i}(u_{h_i}) \approx \phi_{h_{i+1}}(P_i^{i+1}u_{h_i})$. This is not the case for the functions ϕ_{h_i}, which is easy to see when setting $u_{h_i} = 0$ where we obtain a factor 4 for each level. This can be avoided by multiplying ϕ_{h_i} by h_i^2, which leads to the functions

$$\hat{\phi}_{h_i}(u_{h_i}) := \frac{1}{2}u_{h_i}^T A_{h_i} u_{h_i} - \lambda h_i^2 \sum_{j=1}^{n_i} e^{u_{h_i}^j}, \quad i = 1,\dots,r.$$

The construction of a multilevel hierarchy is straightforward. Set $N(i)$ as proposed in Remark 2.1, $\mathcal{V}_i = \mathcal{W}_i = \mathbb{R}^{n_i}$ with the inner product $(\cdot,\cdot)_{h_i}$ and $f_i(x_i, v_i) := \hat{\phi}_{h_i}(x_i + v_i)$ for $i = 1,\dots,r$. The restriction operators are defined by $R_i^{i-1}(x_i, v_i) := I_i^{i-1}(x_i + v_i)$.

2.3. A trust-region algorithm

The algorithm we present in this chapter uses a trust-region framework to ensure global convergence to first-order stationary points. A comprehensive presentation of trust-region algorithms can be found in the monograph [CGT00].

In each iteration, a trust-region method minimizes a simple local model of the objective function around the current iterate. Since the model is assumed to be a good approximation only in a neighborhood of the current iterate, we seek for trial steps that lie inside a *trust region*. The size of this trust region is adaptively controlled by the quality of the model's predictions.

Applied to problem (2.2), in each iteration k on level r a *trial step* $s_{r,k}$ is calculated which is an approximate solution of the *trust-region subproblem*:

$$\min_{s_{r,k}\in\mathcal{V}_r} m_{r,k}(s_{r,k}) \quad \text{subject to } \|s_{r,k}\|_r \le \Delta_{r,k},\ v_{r,k} + s_{r,k} \in C_r, \tag{2.9}$$

where $m_{r,k}$ is a model of the objective function f_r at the current iterate $v_{r,k} \in C_r$, $\|\cdot\|_r$ is a suitable *trust-region norm* and $\Delta_{r,k} > 0$ is the *trust-region radius*. The trial step is required to produce a "sufficient" decrease

$$pred_{r,k} = m_{r,k}(0) - m_{r,k}(s_{r,k})$$

of the model function, which is called *predicted reduction*. Whether the algorithm accepts the step, depends on the ratio $\rho_{r,k}$ between the *actual reduction*

$$ared_{r,k} = f_r(v_{r,k}) - f_r(v_{r,k} + s_{r,k})$$

and its prediction $pred_{r,k}$. If the actual reduction is a sufficiently large fraction of the predicted reduction, i.e., $\rho_{r,k} \geq \eta_1 > 0$, the step is accepted. Otherwise, the size of the trust region was too optimistic and the trust-region radius for the next iteration is decreased by a factor $\gamma_2 < 1$ and the step is rejected. If $\rho_{r,k}$ is close to one, i.e., it satisfies $\rho_{r,k} \geq \eta_2 > \eta_1$, the trust-region radius for the next iteration is increased by a factor $\gamma_1 > 1$.

A common choice for the model function $m_{r,k}$ is the quadratic Taylor approximation of f_r at $v_{r,k}$. This leads to the *quadratic trust-region subproblem*

$$\begin{aligned} &\min_{s_{r,k} \in \mathcal{V}_r} q_{r,k}(s_{r,k}) := \langle g_{r,k}, s_{r,k} \rangle + \frac{1}{2} \langle H_{r,k} s_{r,k}, s_{r,k} \rangle \\ &\text{subject to } \|s_{r,k}\|_r \leq \Delta_{r,k},\ v_{r,k} + s_{r,k} \in C_r, \end{aligned} \tag{2.10}$$

where $g_{r,k} := f_r'(v_{r,k})$ is the first Frï¿½chet derivative of f_r at $v_{r,k}$ and $H_{r,k} \in \mathcal{L}(\mathcal{V}_r, \mathcal{V}_r^*)$ the second-order Gateï¿½ux derivative, or a suitable symmetric approximation of it.

Depending on the structure of the feasible set C_r, there are many well-known algorithms to find good approximate minimizers of the subproblem (2.10), e.g., for problems with simple bounds [CL96, Ulb01]. If, however, the number of unknowns is large, these algorithms become expensive. In the multilevel trust-region algorithm one would like to use the coarser spaces $\mathcal{V}_i$, $i \in N(r)$, and the auxiliary functions f_i by defining a *lower-level trust-region subproblem*

$$\min_{s_i \in \mathcal{V}_i} h_i(s_i) \quad \text{subject to } \|P_i^r s_i\|_r \leq C\Delta_{r,k},\ v_{r,k} + P_i^r s_i \in C_r, \tag{2.11}$$

where h_i is a model of f_r on the space $\mathcal{V}_i$ near the current iterate $v_{r,k}$ and $C > 0$ a constant. Besides the quadratic subproblem, (2.11) can also be used to calculate trial steps if it "is appropriate"[2]. A step $s_{i,*}$ that approximately solves this problem is then prolongated to level r. As in the standard case, the size of the ratio between the reductions of h_i and f_r decides whether the step is accepted and how to change the trust-region radius.

An obvious drawback of (2.11) is the fact that the evaluations of the constraints are made on the finer level, which generally is too expensive. Therefore, we do not use this problem directly. Instead, we relax the constraints such that its evaluation can be solely done on the lower level and the new feasible set is a subset of the feasible set of problem (2.11). This will be discussed in the next sections. Before that we render more precisely what properties the lower-level models h_i must satisfy.

[2]We will discuss a sufficient condition when to use such models in Section 2.3.5

2.3.1. The lower-level model

Contrary to the first conjecture, the function f_i itself in general is not appropriate as model function h_i on the subspace $\mathcal{V}_i$. Without being as general as possible, we will now motivate a condition that must be satisfied by the lower-level models. The idea of trust-region methods is that the agreement between the model and the function increases as the trust-region radius tends to zero. At least, we would expect that in descent directions s_i of the model the fraction $\rho_{r,k}$ of the actual and the predicted reduction tends to one if the length of the step tends to zero, i.e.,

$$\lim_{\substack{\|s_i\|\to 0 \\ \langle h_i'(0),\, s_i\rangle < 0}} \frac{f_r(v_{r,k} + P_i^r s_i) - f_r(v_{r,k})}{h_i(s_i) - h_i(0)} = 1. \tag{2.12}$$

From the assumptions that f_r and h_i are Frï¿½chet differentiable and the prolongation P_i^r is linear and continuous, we obtain

$$\lim_{\substack{\|s_i\|\to 0 \\ \langle h_i'(0),\, s_i\rangle < 0}} \frac{f_r(v_{r,k} + P_i^r s_i) - f_r(v_{r,k})}{h_i(s_i) - h_i(0)} = \lim_{\substack{\|s_i\|\to 0 \\ \langle h_i'(0),\, s_i\rangle < 0}} \frac{\langle f_r'(v_{r,k}), P_i^r s_i\rangle + \mathrm{o}(\|P_i^r s_i\|)}{\langle h_i'(0), s_i\rangle + \mathrm{o}(\|s_i\|)} = \frac{\langle f_r'(v_{r,k}), P_i^r s_i\rangle}{\langle h_i'(0), s_i\rangle}.$$

Hence, a necessary and sufficient condition for (2.12) to hold is $\langle f_r'(v_{r,k}), P_i^r s_i\rangle = \langle h_i'(0), s_i\rangle$ for all $s_i \in \mathcal{V}_i$ with $\langle h_i'(0), s_i\rangle \neq 0$. This leads to the following definition:

Definition 2.1 A continuously differentiable function $h_j \colon \mathcal{V}_j \to \mathbb{R}$ is a *lower-level model* of $h_i \colon \mathcal{V}_i \to \mathbb{R}$ at $v_{i,k}$ if $j \in N(i)$ and

$$\langle h_j'(0), s_j\rangle = \langle h_i'(v_{i,k}), P_j^i s_j\rangle \quad \forall\, s_j \in \mathcal{V}_j. \tag{2.13}$$

Remark 2.2 This condition is slightly stronger than it is necessary to prove global convergence. It would be enough to demand that the error

$$\langle h_j'(0) - (P_j^i)^* h_i'(v_{i,k}), s_j\rangle$$

is small in a certain sense, see, e.g., [CGT00, Section 8.4] for conditions in the standard case. In practice it is no strong restriction to assume (2.13) instead.

A trivial example for a lower-level model of f_r at an iterate $v_{r,k}$ is given by the function $h_i(s_i) := f_r(v_{r,k} + P_i^r s_i)$, which has the obvious disadvantage that its evaluation is in general as expensive as evaluating the original function.

A more reasonable lower-level model of f_r consists of the function f_i and an additional first-order correction term:

$$h_i(s_i) := f_i(x_i, s_i) + \big\langle (P_i^r)^* f_r'(v_{r,k}) - f_i'(x_i, 0), s_i\big\rangle, \tag{2.14}$$

where $x_i := R_r^i(0, v_{r,k})$ is the development point. Besides models of f_r, we also (recursively) need models of models. Assume that h_i is a model on level i and $x_i \in \mathcal{W}_i$ its development

point. Then the first-order corrected model on level $j \in N(i)$ at the point $v_{i,k}$ is given by

$$h_j(s_j) := f_j(x_j, s_j) + \langle (P_j^i)^* h_i'(v_{i,k}) - f_j'(x_j, 0), s_j \rangle, \quad x_j = R_i^j(x_i, v_{i,k}). \tag{2.15}$$

These models are widely used in multilevel optimization methods, for example in [Nas00, GST08, WG09].

Remark 2.3 First-order consistent models are also commonly used in approximation/model management optimization (AMMO) [AL01]. Here, low-fidelity models f_{lo} are used to calculate steps for a high-fidelity model f_{hi} inside an optimization algorithm. To ensure first-order consistency, a modification of f_{lo} similar to (2.15) is often used, the β-correlation approach [CHGK93]. For this, one defines the scaling factor $\beta(s_i) := f_r(v_{r,k} + P_i^r s_i)/f_i(R_r^i(0, v_{r,k}), s_i)$ and builds a local model β_c of β at $s_i = 0$:

$$\beta_c(s_i) = \beta(0) + \langle \nabla\beta(0), s_i \rangle.$$

A straightforward calculation shows that $h_i^\beta(s_i) := \beta_c(s_i) f_i(R_r^i(0, v_{r,k}), s_i)$ is a lower-level model of f_r at $v_{i,k}$ which satisfies (2.13). In comparison to (2.15), these models can only be used when $f_i(R_r^i(0, v_{r,k}), s_i) \neq 0$ holds.

In [GMS$^+$10], lower-level models that are second-order correct were introduced. Besides (2.13), these models also satisfy

$$\langle h_j''(0) s_j, s_j \rangle = \langle h_i''(v_{i,k}) P_j^i s_j, P_j^i s_j \rangle \quad \forall\, s_j \in \mathcal{V}_j. \tag{2.16}$$

By appending an additional second-order correction term to (2.15), we obtain a second-order corrected model of h_i at $v_{i,k}$:

$$\begin{aligned} \bar{h}_j(s_j) := {} & f_j(x_j, s_j) + \langle (P_j^i)^* h_i'(v_{i,k}) - f_j'(x_j, 0), s_j \rangle \\ & + \frac{1}{2} \langle ((P_j^i)^* h_i''(v_{i,k}) P_j^i - f_j''(x_j, 0)) s_j, s_j \rangle, \quad x_j = R_i^j(x_i, v_{i,k}). \end{aligned} \tag{2.17}$$

Remark 2.4 The models (2.15) are also implicitly used in standard nonlinear multigrid methods, e.g., the *Full Approximation Scheme* (FAS) (cf. [Bra77]) or the *Nonlinear Multi-Grid Method* (NMGM) described by Hackbusch in [Hac85, Ch. 9]. For simplicity, we will illustrate the connection only on the basis of a two-grid FAS method, the transfer to more levels is straightforward.

FAS is a method to solve nonlinear systems of the form $L_2(v_2^*) = 0$, where $L_2 \colon \mathbb{R}^{n_2} \to \mathbb{R}^{n_2}$ is the discretization of a nonlinear differential operator. A typical example is the mildly nonlinear operator Φ_{h_2} introduced in Example 2.2. It assumes that a coarser discretization $L_1 \colon \mathbb{R}^{n_1} \to \mathbb{R}^{n_1}$, $n_1 < n_2$, of L_2 and proper prolongation and restriction operators exist. Starting from an iterate $v_{2,0}$ the two-grid iteration consists of two steps:

1. Smoothing: $v_{2,1} = S_2(v_{2,0})$, where S_2 is a smoothing operator, e.g., a nonlinear version of the Gauï¿½-Seidel iteration.

2. Coarse-grid correction: The current iterate is restricted to the coarser grid, $x_1 = R_2^1 v_{2,1}$, and a step v_1^* that (approximately) solves the system

$$L_1(x_1 + v_1) = L_1(x_1) - (P_1^2)^* L_2(v_{2,1}) \tag{2.18}$$

   is calculated. Set $v_{2,2} = v_{2,1} + P_1^2 v_1^*$.

As we have done it here, it is possible to choose different restrictions for the residual $L_2(v_{2,\cdot})$ and the point $v_{2,\cdot}$. Typically, the adjoint of the prolongation operator with respect to a properly scaled Euclidean inner products is used to restrict the residual (cf. Example 2.2).

We will now formulate the FAS method in our (unconstrained) optimization context. For this purpose, we assume that the equations $L_i(v_i^*) = 0$, $i = 1, 2$, can be written as $f_i'(v_i^*) = 0$ where f_i' is the derivative of a function f_i (cf. Example 2.2). The nonlinear Gauï¿½-Seidel step can be formulated as *cyclic coordinate search*: Starting with $i = 1$ and $v_{2,0,0} = v_{2,0}$ we successively seek for all $i = 1, \ldots, n_2$ a minimizer t_i^* of the function $\phi_i(t) := f_2(v_{2,0,i-1} + te_i)$, where e_i is the i-th coordinate direction, and set $v_{2,0,i} = v_{2,0,i-1} + t_i^* e_i$. The iterate $v_{2,1}$ is then set to the resulting vector $v_{2,0,n_2}$. For the coarse-grid correction, we define a lower-level model of the type (2.15) by $h_1(v_1) := f_1(x_1 + v_1) + (v_1, (P_1^2)^* f_2'(v_{2,1}) - f_1'(x_1))$, where $x_1 = R_2^1 v_{2,1}$. A solution v_1^* of the problem

$$\min_{v_1 \in \mathbb{R}^{n_1}} h_1(v_1)$$

satisfies $h_1'(v_1^*) = 0$ and hence

$$f_1'(x_1 + v_1) = f_1'(x_1) - P_2^* f_2'(v_{2,1}), \tag{2.19}$$

which is equivalent to (2.18).

Remark 2.5 If we use the second-order corrected models (2.17) instead of (2.15) in the previous remark, we obtain a different nonlinear multigrid method. A straightforward calculation shows that the resulting algorithm is just the method MNM (Multilevel Nonlinear Method) proposed in [YD06].

2.3.2. The lower-level trust-region subproblem

The lower-level trust-region subproblem (2.11) has some disadvantages that make it hard to solve: On the one hand, both the trust-region and the feasibility condition are evaluated on the space $\mathcal{V}_r$, which is contrary to the effort of using a space with lower dimension for the subproblem. On the other hand, the trust-region condition is not in standard form, which could make it hard to handle.

Therefore, we simplify the subproblem in the following way: First, we introduce level dependent norms $\|\cdot\|_i$ that are compatible with the prolongation operators in the sense that

$$\|\mathcal{P}_j^i s_j\|_i \leq C_{\mathcal{P}} \|s_j\|_j \quad \text{for all } s_j \in C_j \text{ and } j \prec i \tag{2.20}$$

with a level-independent constant $C_{\mathcal{P}} \geq 1$. We call a constant level-independent if it does not depend on the level and does not deteriorate if the number of levels goes to infinity. We replace the first constraint of (2.11) by

$$\|s_i\|_i \leq \Delta_{r,k}.$$

All iterates that satisfy these conditions also satisfy the original trust-region constraint with the constant $C = C_{\mathcal{P}}$.

Second, the constraint

$$v_{r,k} + P_i^r s_i \in C_r$$

of problem (2.11) is replaced by the requirement $s_i \in C_i$, where $C_i \subset \mathcal{V}_i$ is a closed and convex set that satisfies

$$0 \in C_i \quad \text{and} \quad v_{r,k} + P_i^r s_i \in C_r \text{ for all } s_i \in C_i. \tag{2.21}$$

An obvious choice for C_i is the convex set $C_i^{\max}(v_{r,k}) := \{s_i \in \mathcal{V}_i \,|\, v_{r,k} + P_i^r s_i \in C_r\}$, which is just the set used in (2.11). This choice has in general some computational disadvantages: To check whether an element of $\mathcal{V}_i$ is also an element of $C_i^{\max}$, we must prolongate the element and make an evaluation on the fine level, which is expensive. Furthermore, if the set C_r has a special structure, for instance is given by pointwise bounds on the variables, the set $C_i^{\max}$ will in general lose this structure. This is in most cases not desired, because then we have to use a different class of algorithms to solve the trust-region subproblems. We will discuss in Section 4.3 how to construct suitable lower-level sets for the typical case that C_r is given by pointwise bounds.

Summarizing the above, we obtain a simplified lower-level trust-region subproblem

$$\begin{aligned} &\min_{s_i \in \mathcal{V}_i} h_i(s_i) \\ &\text{s.t. } \|s_i\|_i \leq \Delta_{r,k},\ s_i \in C_i, \end{aligned} \tag{2.22}$$

where all evaluations are made on the space $\mathcal{V}_i$. In the following, if we use these subproblems, we call the resulting step a *multilevel step*. Otherwise, if (2.10) was used, we call it a *Taylor* or *smoothing step*.

In general, h_i is a non-quadratic function so that we cannot use standard trust-region subproblem techniques to compute a step for (2.22). However, the problem is similar to (2.2), except for the additional trust-region constraint. Therefore, we calculate steps for (2.22) using the same trust-region method, where we use either a quadratic model of h_i or again recursively a lower-level model of h_i on a level $j \in N(i)$. This is achieved by calling the algorithm on level i with the function h_i, the convex set C_i and by setting the initial trust-region radius $\Delta_{i,0}$ to $\Delta_{r,k}$. In order to ensure the trust-region constraint in (2.22) for the final step, we demand that every successive radius $\Delta_{i,k'}$ satisfies $\Delta_{i,k'} \leq \Delta_{i,0} - \|v_{i,k'} - v_{i,0}\|_i$.

Remark 2.6 Another way of dealing with the additional trust-region constraint is to merge it into C_i by defining the new feasible and convex set $\tilde{C}_i = C_i \cap \{s_i \,|\, \|s_i\|_i \leq \Delta_{i,0}\}$. This was done in [GMTWM08] where problems in $\mathbb{R}^n$ with pointwise bounds were considered and the trust-region norm on every level was given by the maximum norm $\|\cdot\|_\infty$. In this case, the resulting set $\tilde{C}_i$ can also easily be described by pointwise bounds. In general however, the disadvantage of this approach is that if C_i has a special structure, the set $\tilde{C}_i$ will lose it. As an example, consider the case in $\mathbb{R}^n$ of an Euclidean trust-region norm and a box C_i. In particular, this could lead to problems when constructing a new lower-level set C_j, $j \in N(i)$.

Level dependent norms

As outlined in the last section, the simplified lower-level problems use level dependent norms. Because of condition (2.20), they depend on the norm on level r.

If $\mathcal{V}_i$, $i = 1, \dots, r$, are Hilbert spaces with inner product $(\cdot, \cdot)_i$, we can identify the dual space $\mathcal{V}_i^*$ with $\mathcal{V}_i$, which follows from the Riesz representation theorem. In this case we assume that the prolongation operators P_j^i maps from $\mathcal{V}_j$ to $\mathcal{V}_i \cong \mathcal{V}_i^*$ and the adjoint $(P_j^i)^*$ satisfies $(g_i, P_j^i s_j)_i = ((P_j^i)^* g_i, s_j)_j$. In this setting, one can use the norms defined by $\|s_i\|_i := \sqrt{(M_i^r s_i, s_i)_i}$ with the self-adjoint operator $M_i^r := (\mathcal{P}_i^r)^* \mathcal{P}_i^r$. The norm is well defined if $\mathcal{P}_i^r$ is injective and it is easy to see that (2.20) with $C_{\mathcal{P}} = 1$ holds. This type of level dependent norms was first introduced in [GST08] for the special case of the Euclidean norm in $\mathbb{R}^n$. As we will later see in the case of bound constrained programs, our prolongation operator can change in each iteration. This leads to higher computational costs because the operator M_i^r has to be recalculated every time a coarser grid is entered. Even worse, it can happen that the prolongation is not injective and thus the norm is not well defined anymore. In these cases other norms are more suitable.

In a typical multilevel scenario, the spaces $\mathcal{V}_i$ form a nested sequence as for instance in Example 2.1. In this case, the natural prolongation operator is the identity. Hence, every norm on the finest space $\mathcal{V}_r$ could be used as level-dependent trust-region norm. Obviously, (2.20) is satisfied in this case. In Example 2.1 the $H^1(\Omega)$-norm would be a feasible choice.

Let $A \in \mathbb{R}^{n \times m}$ be a matrix. The operator norm $\|A\|_z$, $z \in \{1, 2, \infty\}$, that is induced by the corresponding vector norm $\|\cdot\|_z$ is given by

$$\|A\|_z := \sup_{x \in \mathbb{R}^m} \frac{\|Ax\|_z}{\|x\|_z}.$$

In the setting of Example 2.2, it is easy to see that the operator norm of the prolongation operators satisfies $\|P_i^{i+1}\|_\infty = 1$. Hence, if the maximum-norm is chosen as trust-region norm on each level, (2.20) is valid with $C_{\mathcal{P}} = 1$ because

$$\|\mathcal{P}_i^k s_i\|_k = \|\mathcal{P}_i^k s_i\|_\infty = \|P_{l_m}^k \cdots P_{l_1}^{l_2} P_i^{l_1} s_i\|_\infty \leq \|P_{l_m}^k\|_\infty \cdots \|P_i^{l_1}\|_\infty \|s_i\|_\infty = \|s_i\|_i.$$

The well known inequality $\|A\|_2^2 \leq \|A\|_\infty \|A\|_1$ (see for instance [GVL96, Corollary 2.3.2]) allows us to estimate the Euclidean norm of the prolongation operators by $\|P_i^{i+1}\|_2^2 \leq \|P_i^{i+1}\|_\infty \|P_i^{i+1}\|_1 \leq 4$. This is what we expect considering that $n_{i+1} \approx 4n_i$. If we choose $\|\cdot\|_i = \|\cdot\|_2$ for $i = 1, \dots, r$, assumption (2.20) is satisfied but only with the level dependent constant $C_{\mathcal{P}} = 2^{\sharp r}$. A better choice are the norms induced by the level dependent inner products $(\cdot, \cdot)_{h_i}$, i.e., $\|\cdot\|_i := \sqrt{(\cdot, \cdot)_{h_i}} = h_i \|\cdot\|_2$. They satisfy

$$\|P_i^{i+1} s_i\|_{i+1} \leq h_{i+1} \|P_i^{i+1}\|_2 \|s_i\|_2 \leq h_i \|s_i\|_2 = \|s_i\|_i, \quad i = 1, \dots, r-1,$$

and thus (2.20) with $C_{\mathcal{P}} = 1$.

2.3.3. Stationarity measures

Before we introduce stationarity measures, we recapitulate the first-order necessary optimality condition for the problem

$$\min_{s_i \in C_i} h_i(s_i), \tag{2.23}$$

where C_i is a closed and convex set.

Lemma 2.2 *Assume $h_i\colon C_i \to \mathbb{R}$, $C_i \neq \emptyset$ convex, is a Gï¿½teaux differentiable function and let s_i^* be a local solution of* (2.23)*, then*

$$s_i^* \in C_i \quad \text{and} \quad \langle h_i'(s_i^*), s_i - s_i^* \rangle \geq 0 \quad \text{for all } s_i \in C_i. \tag{2.24}$$

Proof See, for instance [HPUU09, Theorem 1.46].

We call a point s_i^* that satisfies (2.24) a *stationary* or *KKT* point of (2.23).

In this thesis, we use the concept of *stationarity measures* to check for first-order convergence:

Definition 2.2 A continuous function $\chi_i\colon C_i \to \mathbb{R}_+$, C_i convex, is called a *stationarity measure* for problem (2.23) if it satisfies

$$\chi_i(s_i) = 0 \text{ if and only if } s_i \text{ is a KKT-Point of (2.23).} \tag{2.25}$$

In the unconstrained case, i.e., if $C_i = \mathcal{V}_i$, the norm of the derivative is the most commonly used stationarity measure:

$$\chi_i(s_i) = \|h_i'(s_i)\|_{\mathcal{V}_i^*}.$$

Depending on the concrete setting, other choices for the norm are possible.

If $\mathcal{V}_i$ is a Hilbert space, an example of a stationarity measure in the constrained case is the norm of the projected gradient:

$$\chi_i(s_i) := \|s_i - \mathrm{Proj}_{C_i}(s_i - \nabla_{\mathcal{V}_i} h_i(s_i))\|_{\mathcal{V}_i}. \tag{2.26}$$

Here, $\nabla_{\mathcal{V}_i} h_i(s_i)$ is the representation of $h_i'(s_i)$ with respect to the inner product on $\mathcal{V}_i$, i.e., we have the identity

$$\langle h_i'(s_i), v_i \rangle = (\nabla_{\mathcal{V}_i} h_i(s_i), v_i)_{\mathcal{V}_i} \quad \text{for all } v_i \in \mathcal{V}_i.$$

The existence of such an element is just the assertion of the Riesz representation theorem. By

$$\mathrm{Proj}_{C_i}(d_i) = \arg\min_{u_i \in C_i} \|u_i - d_i\|_{\mathcal{V}_i}$$

we denote here the $\mathcal{V}_i$-orthogonal projection of d_i onto C_i.

Another measure mentioned in [CGT00] in the case of $\mathbb{R}^n$, which was also used in conjunction with multgrid optimization in [GMTWM08], is defined by

$$\chi_i^\theta(s_i) := \left| \inf_{\substack{s_i + d_i \in C_i \\ \|d_i\|_{\mathcal{V}_i} \leq \theta}} \langle h_i'(s_i), d_i \rangle \right|, \tag{2.27}$$

where $\theta > 0$ is a fixed constant.

Lemma 2.3 *Let $C_i \subset \mathcal{V}_i$ be a nonempty, closed and convex set and $h_i\colon C_i \to \mathbb{R}$ a continuously differentiable function.*

1. *Let $\mathcal{V}_i$ be a Banach space. The function χ_i^θ defined by* (2.27) *is a stationarity measure.*
2. *Let $\mathcal{V}_i$ be a Hilbert space. Furthermore, let $\nabla_{\mathcal{V}_i} h_i(v)$ be the representation of $h_i'(v) \in \mathcal{V}_i^*$ with respect to the inner product on $\mathcal{V}_i$. The function χ_i defined by* (2.26) *is a stationarity measure.*

PROOF 1. We first show that χ_i^θ is well-defined. We set

$$F_i^\theta(s_i) := \inf_{\substack{s_i + d_i \in C_i \\ \|d_i\|_{\mathcal{V}_i} \leq \theta}} \langle h_i'(s_i), d_i \rangle.$$

For a fixed $s_i \in C_i$, $F_i^\theta(s_i)$ is bounded below by $-\theta \|h_i'(s_i)\|_{\mathcal{V}_i^*}$ because

$$|\langle h_i'(s_i), d_i \rangle| \leq \|d_i\|_{\mathcal{V}_i} \|h_i'(s_i)\|_{\mathcal{V}_i^*} \leq \theta \|h_i'(s_i)\|_{\mathcal{V}_i^*}.$$

Hence, $\chi_i^\theta < \infty$ is satisfied.

Inserting $d_i = 0$ in the definition of F_i shows that $F_i^\theta(s_i) \leq 0$ for all $s_i \in C_i$. This gives $\chi_i^\theta(s_i^*) = 0 \Leftrightarrow \langle h_i'(s_i^*), d_i \rangle \geq 0$ for all d_i with $s_i^* + d_i \in C_i$ and $\|d_i\|_{\mathcal{V}_i} \leq \theta$. Because C_i is convex, this is equivalent to (2.24).

It remains to prove that χ_i^θ is continuous, which is equivalent to the continuity of F_i^θ. Let $s_i \in C_i$ and $\varepsilon > 0$ arbitrary. Since h_i' is continuous we get for $\hat{\varepsilon} = \varepsilon/(2\theta)$, a $\hat{\delta} > 0$ such that $\|h_i'(s_i) - h_i'(\bar{s}_i)\|_{\mathcal{V}_i^*} \leq \hat{\varepsilon}$ and $M \geq 0$ with $\|h_i'(\bar{s}_i)\|_{\mathcal{V}_i^*} \leq M$ for all

$$\bar{s}_i \in B_{\hat{\delta}}(s_i) := \{ s \in \mathcal{V}_i \,|\, \|s - s_i\|_{\mathcal{V}_i} \leq \hat{\delta} \}.$$

Set $\delta = \min\{\hat{\delta}, \varepsilon/(4M)\}$. Let $\bar{s}_i \in B_\delta(s_i) \cap C_i$ and $(d_i^k)_{k \in \mathbb{N}} \subset D(s_i) := \{ d_i \in \mathcal{V}_i \,|\, s_i + d_i \in C_i,\ \|d_i\|_{\mathcal{V}_i} \leq \theta \}$ be a sequence such that $\langle h_i'(s_i), d_i^k \rangle \to F_i^\theta(s_i)$ for $k \to \infty$. For each d_i^k we set $\bar{d}_i^k := \theta/(\delta + \theta)(d_i^k + s_i - \bar{s}_i)$. Note that $\bar{d}_i^k \in D(\bar{s}_i)$ because C_i is convex and $\theta/(\delta + \theta) \leq 1$. We estimate

$$\begin{aligned} |\langle h_i'(s_i), d_i^k \rangle - \langle h_i'(\bar{s}_i), \bar{d}_i^k \rangle| = \Big| & \langle h_i'(s_i) - h_i'(\bar{s}_i), d_i^k \rangle \\ & + (\delta + \theta)^{-1} \left[\delta \langle h_i'(\bar{s}_i), d_i^k \rangle - \theta \langle h_i'(\bar{s}_i), s_i - \bar{s}_i \rangle \right] \Big| \\ \leq \hat{\varepsilon}\theta + M(\delta + \theta)^{-1} 2\delta\theta & \leq \varepsilon/2 + 2M\delta \leq \varepsilon. \end{aligned}$$

Since (d_i^k) is a minimizing sequence and $F_i^\theta(\bar{s}_i) \leq \langle h_i'(\bar{s}_i), d_i^k \rangle$ for all k, it follows that $F_i^\theta(\bar{s}_i) \leq F_i^\theta(s_i) + \varepsilon$. Similar by considering a minimizing sequence $(\bar{d}_i^k)_{k \in \mathbb{N}}$ for $\langle h_i'(\bar{s}_i), \cdot \rangle$ and choosing suitable d_i^k, we obtain $F_i^\theta(s_i) \leq F_i^\theta(\bar{s}_i) + \varepsilon$ with the same ε. This shows $|F_i^\theta(s_i) - F_i^\theta(\bar{s}_i)| \leq \varepsilon$ for all $\bar{s}_i \in B_\delta(s_i) \cap C_i$ and thus the continuity of χ_i^θ.

2. Let $s_i^* \in \mathcal{V}_i$ with $\chi_i(s_i^*) = 0$. From the definition of χ_i follows

$$\chi_i(s_i^*) = 0 \Leftrightarrow \|s_i^* - \mathrm{Proj}_{C_i}(s_i^* - \nabla_{\mathcal{V}_i} h_i(s_i^*))\|_{\mathcal{V}_i} = 0 \Leftrightarrow \mathrm{Proj}_{C_i}(s_i^* - \nabla_{\mathcal{V}_i} h_i(s_i^*)) = s_i^*.$$

Now let $s_i \in C_i$, then with the Projection Theorem A.2 we obtain

$$\langle h_i'(s_i^*), s_i - s_i^* \rangle = (\nabla_{\mathcal{V}_i} h_i(s_i^*), s_i - s_i^*)_{\mathcal{V}_i} = ((s_i^* - \nabla_{\mathcal{V}_i} h_i(s_i^*)) - s_i^*, s_i^* - s_i)_{\mathcal{V}_i} \geq 0.$$

Hence, s_i^* is a KKT-Point. On the other hand if $(s_i^* - \nabla h_i(s_i^*) - s_i^*, s_i^* - s_i)_{\mathcal{V}_i} \geq 0$ for all $s_i \in C_i$, from the alternative definition in the Projection Theorem it follows that $s_i^* = \mathrm{Proj}_{C_i}(s_i^* - \nabla_{\mathcal{V}_i} h_i(s_i^*))$. Hence, $\chi_i(s_i^*) = 0$ if s_i^* is a KKT-Point.

The projection in a Hilbert space on a closed and convex set is continuous (cf. Lemma A.1) and since h_i is continuously differentiable, the continuity of χ_i follows. □

Remark 2.7 If $\mathcal{V}_i$ is a reflexive Banach space, then for every $s_i \in \mathcal{V}_i$ exists $d_i^* \in C_i$ with $\|d_i^*\|_{\mathcal{V}_i} \leq \theta$ that realizes the minimum of (2.27), i.e.,

$$\langle h_i'(s_i), d_i^* \rangle = \min_{\substack{s_i + d_i \in C_i \\ \|d_i\|_{\mathcal{V}_i} \leq \theta}} \langle h_i'(s_i), d_i \rangle.$$

This result is a straightforward conclusion of Theorem A.3 and Lemma A.2.

The next lemma shows that if the projected gradient is well defined, there is a correlation between both stationary measures.

Lemma 2.4 *Let $\mathcal{V}_i$ be a Hilbert space. Under the assumptions of Lemma 2.3 2., the projected gradient $p(s_i) := Proj_{C_i}(s_i - \nabla_{\mathcal{V}_i} h_i(s_i)) - s_i$ is a solution of*

$$\min_{\substack{s_i + d_i \in C_i \\ \|d_i\|_{\mathcal{V}_i} \leq \theta}} \langle h_i'(s_i), d_i \rangle = \min_{\substack{s_i + d_i \in C_i \\ \|d_i\|_{\mathcal{V}_i} \leq \theta}} (\nabla_{\mathcal{V}_i} h_i(s_i), d_i)_{\mathcal{V}_i} \tag{2.28}$$

with $\theta = \|p(s_i)\|_{\mathcal{V}_i}$.

Proof In the following we use the set $D_i := \{d_i \in \mathcal{V}_i \,|\, s_i + d_i \in C_i,\ \|d_i\|_{\mathcal{V}_i} \leq \theta\}$. If $d_i^* \in D_i$ is a solution of (2.28), then

$$(\nabla_{\mathcal{V}_i} h_i(s_i), d_i^*)_{\mathcal{V}_i} \leq (\nabla_{\mathcal{V}_i} h_i(s_i), d_i)_{\mathcal{V}_i} \Leftrightarrow (-\nabla_{\mathcal{V}_i} h_i(s_i),\, d_i^* - d_i)_{\mathcal{V}_i} \geq 0 \quad \text{for all } d_i \in D_i.$$

Let $\bar{d}_i \in D_i$ be an element with $\|\bar{d}_i\|_{\mathcal{V}_i} = \theta$. For every $d_i \in D_i$,

$$(-\nabla_{\mathcal{V}_i} h_i(s_i),\, \bar{d}_i - d_i)_{\mathcal{V}_i} = (-\nabla_{\mathcal{V}_i} h_i(s_i) - \bar{d}_i,\, \bar{d}_i - d_i)_{\mathcal{V}_i} + (\bar{d}_i,\, \bar{d}_i - d_i)_{\mathcal{V}_i}$$

holds. Since

$$0 \leq \|\bar{d}_i - d_i\|_{\mathcal{V}_i}^2 = \|\bar{d}_i\|_{\mathcal{V}_i}^2 + \|d_i\|_{\mathcal{V}_i}^2 - 2(\bar{d}_i, d_i)_{\mathcal{V}_i} \Rightarrow (\bar{d}_i, d_i)_{\mathcal{V}_i} \leq \theta^2,$$

it follows that $(-\nabla_{\mathcal{V}_i} h_i(s_i),\, \bar{d}_i - d_i)_{\mathcal{V}_i} \geq (-\nabla_{\mathcal{V}_i} h_i(s_i) - \bar{d}_i,\, \bar{d}_i - d_i)_{\mathcal{V}_i}$. Setting $\bar{d}_i = p(s_i)$ and using the Projection Theorem (A.2) yields

$$(s_i - \nabla_{\mathcal{V}_i} h_i(s_i) - \mathrm{Proj}_{C_i}(s_i - \nabla_{\mathcal{V}_i} h_i(s_i)), \mathrm{Proj}_{C_i}(s_i - \nabla_{\mathcal{V}_i} h_i(s_i)) - (s_i + d_i))_{\mathcal{V}_i} \geq 0 \quad \text{for all } s_i + d_i \in C_i.$$

Hence, we get $(-\nabla_{\mathcal{V}_i} h_i(s_i),\, p(s_i) - d_i)_{\mathcal{V}_i} \geq 0$ for all $d_i \in D_i$, which shows that $p(s_i)$ is a solution of (2.28). □

Both stationarity measure can be quite expensive to evaluate depending on the space $\mathcal{V}_i$. In our typical setting, where $\mathcal{V}_i$ is a finite dimensional subset of $H^1(\Omega)$, the computation of the projected gradient involves the calculation of a representation and the projection with respect to the inner product on $H^1(\Omega)$, which is very expensive. In Chapter 4, we will therefore consider a typical multilevel setting and introduce a multilevel stationarity measure which is well suited and could be evaluated relatively cheap in a concrete implementation.

2.3.4. Cauchy decrease condition

A trust-region algorithm is expected to converge to a local solution only if the trial steps produce a sufficiently large decrease of the model function. A well-establish way to impose such a condition is the requirement that the decrease provided by the trial step should be at least a fraction of the *Cauchy decrease*. In the unconstrained case, the Cauchy decrease denotes the maximum model reduction along the steepest descent direction of the trust-region subproblem. We impose the following *fraction of Cauchy decrease condition* for every Taylor step $s_{i,k}$ in our algorithm:

$$pred_{i,k} = -q_{i,k}(s_{i,k}) \geq \kappa_{\text{mdc}}\, \chi_i(v_{i,k}) \min\left[1, \frac{\chi_i(v_{i,k})}{\beta_C}, \Delta_{i,k}\right]. \tag{2.29}$$

with constants $\kappa_{\text{mdc}} > 0$ and $\beta_C \geq 1$. One of our goals in the construction of the algorithm is the level-independence in examples like Example 2.1 or 2.2. For this it is necessary that the constants which appear in the condition must not depend on the level and the mesh-size of the discretization. In Chapters 3 and 4 we will analyse various algorithms that approximately solve the trust-region subproblems, which satisfy (2.29).

We will see in the convergence proof of the trust-region method that a condition similar to (2.29) with different constants also automatically holds for the multilevel steps in our algorithm.

2.3.5. Smoothness property

In classical multigrid theory, the usage of coarser grids is only reasonable if the error is smooth enough. A similar problem occurs for the multilevel step. From the definition of the lower-level models follows for the derivative at the origin of a model h_j

$$h_j'(0) = (P_j^i)^* h_i'(v_{i,k}).$$

In most applications, the kernel of $(P_j^i)^*$ is much larger than its range. In Example 2.4 the prolongation operators P_i^{i+1} map from $\mathbb{R}^{n_i} \to \mathbb{R}^{n_{i+1}}$ where $4n_i \approx n_{i+1}$. The prolongation is injective and hence from $\ker((P_i^{i+1})^*) = \text{image}(P_i^{i+1})^\perp$ follows $\dim(\ker((P_i^{i+1})^*)) \approx 3n_i$. So it is possible that the origin is already a (nearly) stationary point of the lower-level model h_j. In this case, we cannot expect a good step that produces a reasonable descent of the lower-level model.

A similar problem can occur if the feasible set C_j of the simplified lower level problem (2.22) is too small compared to C_i. This depends of course on the construction of the lower-level set, but even in the case $C_j = C_j^{\max}$ the set could be equal to $\{0\}$. Consider as an example for this Figure 2.5, where on level 2 the set $C_2 = [l_2, u_2]$, the shaded area, consists of pointwise bounds on the steps. As prolongation we use standard linear interpolation. There are non-zero steps in this set, e.g., the step s_2 as shown in the figure. But on the lower level, every step in $C_1^{\max}$ must be equal to zero, because otherwise it would violate either the lower or the upper bound at the nodes that are also on the coarse grid.

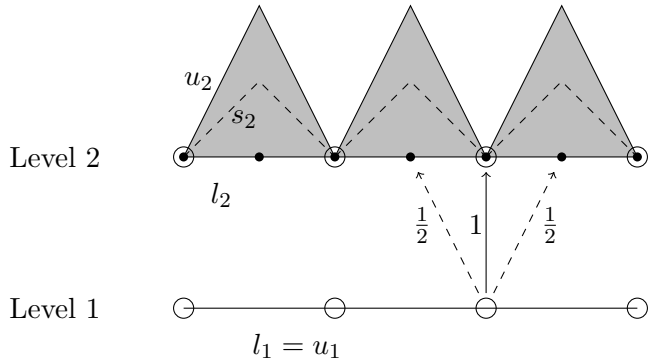

Figure 2.5.: Example of a zero set $C_1^{\max}$ on level 1

These considerations motivates that we only should use the lower-level models on level $j \in N(i)$ when its origin is not "too stationary" in comparison to the current iterate. Indeed, it turns out that a sufficient condition, which guarantees an adequate descent of the multilevel step, is the following *smoothness property*:

$$\chi_j(0) \geq \kappa_\chi \chi_i(v_{i,k}), \quad 0 < \kappa_\chi \leq 1. \tag{2.30}$$

When this condition is not satisfied, we make a Taylor iteration. In comparison to usual trust-region methods we will not try to solve the trust-region subproblem as good as possible. Instead, we use a cheaper algorithm that has a smoothing effect such that (2.30) is more likely to be fulfilled in the next iteration. Of course these steps must satisfies the Cauchy decrease condition. We will see in Chapter 3 and Chapter 4 that the violation of the smoothness property is important to show (2.29) for the smoothing steps. The situation is different for Taylor steps on the coarsest levels where $N(i) = \emptyset$. In this case we use a standard algorithm to obtain a step which approximately solves the trust-region subproblem.

2.3.6. The algorithm TRMLConv

After these preliminaries, we formulate the complete algorithm:

Algorithm 2.1 (TRMLConv(i, h_i, $\Delta_{i,0}$, $\hat{x}_i$, C_i))
Choose $0 < \eta_1 < \eta_2 \leq 1,\ \gamma_1 > 1,\ \gamma_2 < 1,\ \kappa_\chi \in (0, 1]$ and

$$\varepsilon_i^\chi > 0,\ 0 < \varepsilon_i^\Delta < 1 \quad \text{for } i = 1, \ldots, r.$$

Step 0: Initialization
Set $k = 0$. If $i = r$, set $v_{r,0} = \hat{x}_r$ and $x_r = 0$, otherwise set $v_{i,0} = 0$ and $x_i = \hat{x}_i$.

Step 1: Model choice
If $N(i) = \emptyset$ (coarsest level), go to Step 3 (Taylor step). If (2.30) and

$$\chi_j(0) \geq \varepsilon_j^\chi \tag{2.31}$$

are satisfied for at least one $j \in N(i)$, go to Step 2 (Multilevel step) or Step 3. Otherwise, go to Step 3.

Step 2: Multilevel step computation
Choose $j \in N(i)$ and define a lower-level model h_j of h_i at $v_{i,k}$, such that (2.30) and (2.31) are satisfied. Furthermore, determine a transfer operator $P_j^i \colon \mathcal{V}_j \to \mathcal{V}_i$ and a convex set C_j such that conditions (2.21) are satisfied. Call

$$\text{TRMLConv}(j,\, h_j, \Delta_{i,k},\, R_i^j(x_i, v_{i,k}),\, C_j)$$

which returns with a step $v_{j,*}$.
Set $s_{i,k} = P_j^i v_{j,*}$ and $pred_{i,k} = h_j(0) - h_j(v_{j,*})$. Go to Step 4.

Step 3: Taylor step computation
Choose an approximation $H_{i,k} \in \mathcal{L}(\mathcal{V}_i, \mathcal{V}_i^*)$ of $h_i''(v_{i,k})$. Compute an approximate solution $s_{i,k}$ of the trust-region subproblem

$$\begin{aligned} \min_{s_{i,k} \in \mathcal{V}_i} q_{i,k}(s_{i,k}) &:= \langle h_i'(v_{i,k}), s_{i,k} \rangle + \frac{1}{2} \langle H_{i,k} s_{i,k}, s_{i,k} \rangle \\ \text{subject to } \|s_{i,k}\|_i \leq \Delta_{i,k},&\ \ v_{i,k} + s_{i,k} \in C_i, \end{aligned} \tag{2.32}$$

that satisfies the fraction of Cauchy decrease condition (2.29). Set $pred_{i,k} = -q_{i,k}(s_{i,k})$.

Step 4: Acceptance of the trial point
Set $ared_{i,k} = h_i(v_{i,k}) - h_i(v_{i,k} + s_{i,k})$ and $\rho_{i,k} = ared_{i,k}/pred_{i,k}$.
If $\rho_{i,k} \geq \eta_1$, set $v_{i,k+1} = v_{i,k} + s_{i,k}$, otherwise set $v_{i,k+1} = v_{i,k}$. Define

$$\Delta_{i,k}^+ := \begin{cases} \gamma_1 \Delta_{i,k} & \text{if } \rho_{i,k} \geq \eta_2, \\ \Delta_{i,k} & \text{if } \eta_1 \leq \rho_{i,k} < \eta_2, \\ \gamma_2 \Delta_{i,k} & \text{if } \rho_{i,k} < \eta_1, \end{cases} \tag{2.33}$$

and set

$$\Delta_{i,k+1} = \begin{cases} \min\{\Delta_{i,k}^+, \Delta_{i,0} - \|v_{i,k+1}\|_i\} & \text{if } i < r, \\ \Delta_{i,k}^+ & \text{if } i = r. \end{cases} \tag{2.34}$$

Step 5: Termination
If $\chi_i(v_{i,k+1}) \leq \varepsilon_i^\chi$ or if $i < r$ and

$$\|v_{i,k+1}\|_i > (1 - \varepsilon_i^\Delta)\Delta_{i,0}, \tag{2.35}$$

return with $v_{i,k+1}$. Otherwise, set $k \leftarrow k + 1$ and go to Step 1.
One is also free to terminate if $i < r$ and at least one successful step was already made.

The algorithm on level r is started by calling TRMLConv(r, f_r, $\Delta_{r,0}$, $\hat{x}_r$, C_r), where $\Delta_{r,0}$ is the initial trust-region radius and $\hat{x}_r$ the initial point of the algorithm.

Remark 2.8 For the evaluation of (2.30) in Step 1 we actually have to construct the lower-level model h_j and the convex set C_j, which must be the same as in Step 2 of the algorithm.

Remark 2.9 If we make a multilevel step at level i in iteration k and enter level $j \in N(i)$, the initial trust-region radius $\Delta_{j,0}$ satisfies $\Delta_{j,0} = \Delta_{i,k}$. This fact will often be used in the following.

Remark 2.10 The trust-region update rule (2.33) can be altered in various ways without changing the global convergence properties of the algorithm. For instance we could allow the following, more general update rule:

$$\text{Choose } \Delta_{i,k}^{+} \in \begin{cases} (\Delta_{i,k}, \gamma_1 \Delta_{i,k}] & \text{if } \rho_{i,k} \geq \eta_2, \\ [\gamma_2 \Delta_{i,k}, \Delta_{i,k}] & \text{if } \eta_1 \leq \rho_{i,k} < \eta_2, \\ (\gamma_3 \Delta_{i,k}, \gamma_2 \Delta_{i,k}] & \text{if } \rho_{i,k} < \eta_1, \end{cases}$$

with an additional constant $\gamma_3 < \gamma_2$.

Remark 2.11 Condition (2.31) ensures that we have to make at least one successful step on the coarser level before the algorithm terminates.

In the following, we call an iteration (i, k) *successful* (*very successful*) if $\rho_{i,k} \geq \eta_1$ ($\rho_{i,k} \geq \eta_2$) in Step 4 of the algorithm, otherwise we call it *unsuccessful*.

2.4. Global convergence

The proof of global convergence follows the classical proofs of trust-region methods, but the methods are more technical. On the one hand, this is because of the multilevel setting, on the other hand it comes from the need to obtain estimates that are independent from constants that become worse as the number of levels increases. One example is the norm of the Hessian matrices of the fine level function. In the classical theory it is common to demand that these norms are bounded by a constant that occurs in many places of the proof. For multilevel optimization problems like Example 2.1 the discrete L^2-norm of the Hessians is of size $\mathcal{O}(h^{-2})$ where h is equal to the mesh size. As we will see later, this is also important for the choices of the stationarity measure and the level dependent trust-region norms.

The first lemma shows that a step generated by Algorithm 2.1 violates the trust-region condition at most by the factor $C_{\mathcal{P}}$ from (2.20).

Lemma 2.5 *Let the trust-region norms $\|\cdot\|_i$ satisfy* (2.20) *and let $s_{i,k}$ be generated by Step 2 or Step 3 of Algorithm 2.1. Then $\|s_{i,k}\|_i \leq C_{\mathcal{P}} \Delta_{i,k}$ holds.*

Proof If $s_{i,k}$ is generated by Step 2 of the algorithm, the assumption follows directly from (2.22). Hence, in the following we assume that (i, k) is a multilevel iteration on level $j \in N(i)$ and $s_{i,k} = P_j^i v_{j,*}$. Without loss of generality, we assume that iteration $(* - 1)$ is the last successful

iteration on each level. Therefore, $s_{i,k} = P_j^i v_{j,*} = P_j^i(v_{j,*-1} + s_{j,*-1})$. If $s_{j,*-1}$ is a Taylor step, we obtain

$$\|s_{i,k}\|_i \leq C_{\mathcal{P}} \|v_{j,*-1} + s_{j,*-1}\|_j \leq C_{\mathcal{P}}(\|v_{j,*-1}\|_j + \Delta_{j,*-1}).$$

From (2.34) it follows that $\Delta_{j,*-1} \leq \Delta_{j,0} - \|v_{j,*-1}\|_j$ and thus $\|s_{i,k}\|_i \leq C_{\mathcal{P}} \Delta_{i,k}$.

If instead $s_{j,*-1}$ is a multilevel step, we further decompose the iteration until we reach a level l_m, $l_m \prec l_1 = j$, where the last successful step was a Taylor step. We get

$$\begin{aligned} s_{i,k} &= P_{l_1}^i (v_{l_1,*-1} + P_{l_2}^{l_1}(v_{l_2,*-1} + P_{l_3}^{l_2}(\ldots + P_{l_m}^{l_{m-1}}(v_{l_m,*-1} + s_{l_m,*-1})\ldots))) \\ &= \sum_{k=1}^{m} \mathcal{P}_{l_k}^i v_{l_k,*-1} + \mathcal{P}_{l_m}^i s_{l_m,*-1}. \end{aligned}$$

With (2.20) follows

$$\begin{aligned} \|s_{i,k}\|_i &\leq \sum_{k=1}^{m} \|\mathcal{P}_{l_k}^i v_{l_k,*-1}\|_i + \|\mathcal{P}_{l_m}^i s_{l_m,*-1}\|_i \leq C_{\mathcal{P}} \sum_{k=1}^{m} (\|v_{l_k,*-1}\|_{l_k} + \|s_{l_m,*-1}\|_{l_m}) \\ &\leq C_{\mathcal{P}} \sum_{k=1}^{m} (\|v_{l_k,*-1}\|_{l_k} + \Delta_{l_m,*-1}). \end{aligned}$$

Repeated application of (2.34) for the iteration $* - 1$ on levels $l_m, l_{m-1}, \ldots, l_1$ yields

$$\begin{aligned} \|s_{i,k}\|_i &\leq C_{\mathcal{P}} \sum_{k=1}^{m-1} (\|v_{l_k,*-1}\|_{l_k} + \Delta_{l_m,0}) = C_{\mathcal{P}} \sum_{k=1}^{m-1} (\|v_{l_k,*-1}\|_{l_k} + \Delta_{l_{m-1},*-1}) \\ &\leq \ldots = C_{\mathcal{P}}(\|v_{j,*-1}\|_j + \Delta_{j,*-1}) \leq C_{\mathcal{P}} \Delta_{i,k}. \end{aligned}$$

□

Corollary 2.1 *All iterates $v_{j,k}$ with $j < r$ generated by Algorithm 2.1 satisfy $\|v_{j,k}\|_j \leq C_{\mathcal{P}} \Delta_{j,0}$. In particular, if $s_{i,k} = \mathcal{P}_j^i v_{j,*}$ is a multilevel step, $\|v_{j,*}\|_j \leq C_{\mathcal{P}} \Delta_{i,k}$ holds.*

Proof Since $v_{j,0} = 0$, the assertion is true for $k = 0$. Hence, we assume $k > 0$. Using the previous lemma, (2.34) and $C_{\mathcal{P}} \geq 1$ we conclude

$$\begin{aligned} \|v_{j,k}\|_j &\leq \|v_{j,k-1}\|_j + \|s_{j,k-1}\|_j \leq \|v_{j,k-1}\|_j + C_{\mathcal{P}} \Delta_{j,k-1} \\ &\leq (1 - C_{\mathcal{P}}) \|v_{j,k-1}\|_j + C_{\mathcal{P}} \Delta_{j,0} \leq C_{\mathcal{P}} \Delta_{j,0}. \end{aligned}$$

The second statement now follows directly from Remark 2.9. □

For the global convergence theory we need further assumptions on the lower-level model functions h_i. First of all, we assume that h_i possesses the same differentiability properties as the functions $v_i \mapsto f_i(x_i, v_i)$. This means that all models h_i are continuously differentiable and that the second-order Gateï¿½ux derivatives exist and the mappings $v_i \mapsto h_i''(v_i)[s, s]$ are continuous for all directions $s \in \mathcal{V}_i$. This is obviously satisfied for the first- and second-order corrected models (2.15) and (2.17).

The other assumptions concern the approximation of the Hessian used in the quadratic model. We assume that there exists a constant $\beta_1 \geq 0$ such that for all $i \in \{1, \ldots, r\}$, iterates $v_{i,k} \in C_i$, feasible steps $s_{i,k}$ and $t \in [0,1]$

$$\left|\left\langle (H_{i,k} - h_i''(v_{i,k} + ts_{i,k}))s_{i,k}, s_{i,k}\right\rangle\right| \leq 2\beta_1 \|s_{i,k}\|_i^2 \tag{2.36a}$$

is satisfied, where $H_{i,k}$ is the approximation used in the quadratic trust-region subproblem (2.32) at the point $v_{i,k}$. Note that from the definition of the algorithm $h_r = f_r$ follows. The second assumption is needed for the multilevel step and demands that for all i with $N(i) \neq \emptyset$ and all k the Hessians of the lower-level models h_j of h_i at $v_{i,k}$ are related in the sense that for all $v_j \in C_j$ and $t \in [0,1]$

$$\left|\left\langle (h_j''(tv_j) - (P_j^i)^* h_i''(v_{i,k} + tP_j^i v_j)P_j^i)v_j, v_j\right\rangle\right| \leq 2\beta_2 \|v_j\|_j^2 \tag{2.36b}$$

holds.

Remark 2.12 If (2.36a) is satisfied for $H_{i,k} = h_i''(v_{i,k})$ and

$$\left|\left\langle (h_j''(0) - (P_j^i)^* h_i''(v_{i,k})P_j^i)v_j, v_j\right\rangle\right| \leq C\|v_j\|_j^2 \quad \text{for all } v_j \in C_j \tag{2.37}$$

holds for all iterates $v_{i,k}$, assumption (2.36b) is also satisfied:

$$\begin{aligned}
&\left|\left\langle (h_j''(tv_j) - (P_j^i)^* h_i''(v_{i,k} + tP_j^i v_j)P_j^i)v_j, v_j\right\rangle\right| \\
&\quad \leq \left|\left\langle (h_j''(tv_j) - (P_j^i)^* h_i''(v_{i,k})P_j^i)v_j, v_j\right\rangle\right| + \left|\left\langle (P_j^i)^*\big(h_i''(v_{i,k}) - h_i''(v_{i,k} + tP_j^i v_j)\big)P_j^i v_j, v_j\right\rangle\right| \\
&\quad \leq \left|\left\langle (h_j''(tv_j) - h_j''(0))v_j, v_j\right\rangle\right| + \left|\left\langle (h_j''(0) - (P_j^i)^* h_i''(v_{i,k})P_j^i)v_j, v_j\right\rangle\right| + 2\beta_1\|P_j^i v_j\|_i^2 \\
&\quad \leq 2\beta_1\|v_j\|_j^2 + C\|v_j\|_j^2 + 2\beta_1 C_{\mathcal{P}}^2\|v_j\|_j^2 \leq (2\beta_1(1 + C_{\mathcal{P}}^2) + C)\|v_j\|_j^2.
\end{aligned}$$

This shows (2.36b) with $\beta_2 = 2\beta_1(1 + C_{\mathcal{P}}^2) + C$. In the case of second-order corrected models, e.g., when using the model defined by (2.17), assumption (2.37) is satisfied with $C = 0$, which follows directly from (2.16).

The last assumption on the models is only needed to ensure that the algorithm terminates after a finite amount of time and is always satisfied if the spaces $\mathcal{V}_i$ are finite dimensional, which is the typical case. The models h_i must be bounded below on every ball $B_\Delta(0) := \{v_i \in \mathcal{V}_i \mid \|v_i\|_i \leq \Delta\}$ with $0 < \Delta < \infty$. If $\mathcal{V}_i$ is infinite dimensional, this must not necessarily be true since balls are not compact. However, even in this case the assumption can be shown for the first- and second-order corrected models if all functions f_i are bounded below and the trust-region norms $\|\cdot\|_i$ satisfies $\|v_i\|_{\mathcal{V}_i} \leq C\|v_i\|_i$ with a fixed constant $C > 0$. Let h_j, $j \in N(i)$, be a second-order corrected model of h_i at v_i. Since f_j and h_i are twice Gï¿½teaux differentiable, we can estimate

$$\begin{aligned}
h_j(v_j) &= f_j(x_j, v_j) + \langle (P_j^i)^* h_i'(v_i) - f_j'(x_j, 0), v_j\rangle + \frac{1}{2}\langle ((P_j^i)^* h_i''(v_i)P_j^i - f_j''(x_j, 0))s_j, s_j\rangle \\
&\geq f_j(x_j, v_j) - \|(P_j^i)^* h_i'(v_i) - f_j'(x_j, 0)\|_{\mathcal{V}_i^*}\|v_j\|_{\mathcal{V}_i} - \|(P_j^i)^* h_i''(v_i)P_j^i - f_j''(x_j, 0)\|_{\mathcal{L}(\mathcal{V}_i, \mathcal{V}_i^*)}\|v_j\|_{\mathcal{V}_i}^2 \\
&\geq f_j(x_j, v_j) - C(v_i, x_j)\max\{1, \Delta^2\},
\end{aligned}$$

where $C(v_i, x_j)$ is a constant that does not depend on v_j. Since f_j is bounded below, this shows the assertion for second-order corrected models. The argumentation for the first-order corrected models is nearly identical.

For the upcoming results, we generally assume that all lower-level models used in Algorithm 2.1 satisfy (2.36a) and (2.36b).

Lemma 2.6 *The estimate*

$$|pred_{i,k} - ared_{i,k}| \leq \beta \Delta_{i,k}^2$$

with $\beta := C_{\mathcal{P}}^2 \max\{1, \beta_1, \beta_2\}$ *holds in every iteration of Algorithm 2.1.*

Proof We have to distinguish whether $s_{i,k}$ is a multilevel or a smoothing step. Suppose $s_{i,k}$ was generated by Step 2. Then the predicted reduction $pred_{i,k}$ is equal to

$$\begin{aligned} pred_{i,k} &= -q_{i,k}(s_{i,k}) = -\langle h_i'(v_{i,k}), s_{i,k}\rangle - \frac{1}{2}\langle H_{i,k}s_{i,k}, s_{i,k}\rangle \\ &= -\langle h_i'(v_{i,k}), s_{i,k}\rangle - \int_0^1 (1-t)\langle H_{i,k}s_{i,k}, s_{i,k}\rangle \,\mathrm{d}t. \end{aligned}$$

By Taylor's Theorem with integral remainder term (cf. Lemma A.3), we obtain for the actual reduction

$$ared_{i,k} = h_i(v_{i,k}) - h_i(v_{i,k} + s_{i,k}) = -\langle h_i'(v_{i,k}), s_{i,k}\rangle - \int_0^1 (1-t)\langle h_i''(v_{i,k} + ts_{i,k})s_{i,k}, s_{i,k}\rangle \,\mathrm{d}t.$$

With assumption (2.36a) and Lemma 2.5, the rest follows straightforward:

$$\begin{aligned} |pred_{i,k} - ared_{i,k}| &= \left| \int_0^1 (1-t)\langle (H_{i,k} - h_i''(v_{i,k} + ts_{i,k}))s_{i,k}, s_{i,k}\rangle \,\mathrm{d}t \right| \\ &\leq 2\int_0^1 (1-t)\beta_1 \|s_{i,k}\|_i^2 \,\mathrm{d}t \leq \beta_1 \|s_{i,k}\|_i^2 \\ &\leq \beta_1 \|s_{i,k}\|_i^2 \leq \beta_1 C_{\mathcal{P}}^2 \Delta_{i,k}^2. \end{aligned}$$

Let us now consider the case where $s_{i,k} = P_j^i v_{j,*}$ is a multilevel step. We use Taylor's Theorem for both the actual and the predicted reduction:

$$\begin{aligned} ared_{i,k} &= h_i(v_{i,k}) - h_i(v_{i,k} + P_j^i v_{j,*}) \\ &= -\langle h_i'(v_{i,k}), P_j^i v_{j,*}\rangle - \int_0^1 (1-t)\langle h_i''(v_{i,k} + tP_j^i v_{j,*})P_j^i v_{j,*}, P_j^i v_{j,*}\rangle \,\mathrm{d}t, \\ pred_{i,k} &= h_j(0) - h_j(v_{j,*}) \\ &= -\langle h_j'(0), v_{j,*}\rangle - \int_0^1 (1-t)\langle h_j''(tv_{j,*})v_{j,*}, v_{j,*}\rangle \,\mathrm{d}t. \end{aligned}$$

From the definition of the lower-level models (2.13), it follows that $\langle h_j'(0), v_{j,*}\rangle = \langle h_i'(v_{i,k}), P_j^i v_{j,*}\rangle$. Thus, we get for the difference

$$|pred_{i,k} - ared_{i,k}| = \left| \int_0^1 (1-t)\langle (h_j''(tv_{j,*}) - (P_j^i)^* h_i''(v_{i,k} + tP_j^i v_{j,*})P_j^i)v_{j,*}, v_{j,*}\rangle \,\mathrm{d}t \right|.$$

Using (2.36b) and Corollary 2.1 we get by the same argument as in the first case:

$$|pred_{i,k} - ared_{i,k}| \leq \beta_2 \|v_{j,*}\|_j^2 \leq \beta_2 C_{\mathcal{P}}^2 \Delta_{i,k}^2.$$

Taking the maximum of the estimates finishes the proof. □

The previous lemma shows that the prediction error between a function h_i and its model decreases at least quadratically with the size of the trust-region. This holds in both cases if we use the quadratic approximation and the multilevel model, where for the latter property (2.13) is essential.

Remark 2.13 For the proof of the global convergence we are only interested in the difference of the reductions for "small" steps, i.e., how the models behave locally. Therefore, it is enough to demand that (2.36a) and (2.36b) hold for steps $s_{i,k}$ resp. v_j whose norms are bounded by a fixed positive constant.

The next lemma shows that every step of our algorithm is very successful, whenever the trust region is small enough.

Lemma 2.7 *Let $s_{i,k}$ be a step generated by Algorithm 2.1. Iteration (i,k) is very successful and*

$$ared_{i,k} = h_i(v_{i,k}) - h_i(v_{i,k} + s_{i,k}) \geq \eta_2^{\sharp i+1} \kappa_{mdc}\, \kappa_\chi^{\sharp i} \chi_i(v_{i,k}) \Delta_{i,k} \tag{2.38}$$

holds whenever

$$\Delta_{i,k} \leq \min\left\{1, \kappa_{mdc}\, \frac{\eta_2^{\sharp i} \kappa_\chi^{\sharp i} \chi_i(v_{i,k})(1-\eta_2)}{\beta}, \frac{\kappa_\chi^{\sharp i} \chi_i(v_{i,k})}{\beta_C}\right\}. \tag{2.39}$$

Proof We first consider the case where $s_{i,k}$ is a Taylor-step. It satisfies the fraction of Cauchy decrease condition (2.29) and because $\Delta_{i,k} \leq \min\{1,\, \chi_i(v_{i,k})/\beta_C\}$ we obtain for the predicted reduction

$$pred_{i,k} = -q_{i,k}(s_{i,k}) \geq \kappa_{\mathrm{mdc}}\, \chi_i(v_{i,k}) \Delta_{i,k}.$$

Using Lemma 2.6, (2.39) and $\eta_2, \kappa_\chi \leq 1$ we estimate

$$\frac{pred_{i,k} - ared_{i,k}}{pred_{i,k}} \leq \frac{\beta \Delta_{i,k}^2}{\kappa_{\mathrm{mdc}}\, \chi_i(v_{i,k}) \Delta_{i,k}} \leq \kappa_\chi^{\sharp i} \eta_2^{\sharp i} (1-\eta_2) \leq (1-\eta_2),$$

which leads to

$$\rho_{i,k} = \frac{ared_{i,k}}{pred_{i,k}} \geq \eta_2.$$

Therefore, the step is very successful and

$$h_i(v_{i,k}) - h_i(v_{i,k} + s_{i,k}) \geq -\eta_2 q_{i,k}(s_{i,k}) \geq \eta_2 \kappa_{\mathrm{mdc}}\, \chi_i(v_{i,k}) \Delta_{i,k} \geq \eta_2^{\sharp i+1} \kappa_{\mathrm{mdc}}\, \kappa_\chi^{\sharp i} \chi_i(v_{i,k}) \Delta_{i,k}.$$

We use induction to prove the multilevel case. Note that at the latest on levels l with $N(l) = \emptyset$, we have to make Taylor steps for which the lemma was already proven. So in the following, we assume that the statement of the lemma holds on level $j \in N(i)$, which was entered in iteration (i,k).

In this case, the smoothness property (2.30) is satisfied for j. Thus, by assumption (2.39) follows

$$\begin{aligned}
\Delta_{j,0} = \Delta_{i,k} &\leq \min\left\{1, \kappa_{\mathrm{mdc}}\, \frac{\eta_2^{\sharp i} \kappa_\chi^{\sharp i} \chi_i(v_{i,k})(1-\eta_2)}{\beta}, \frac{\kappa_\chi^{\sharp i} \chi_i(v_{i,k})}{\beta_C}\right\} \\
&\leq \min\left\{1, \kappa_{\mathrm{mdc}}\, \frac{\eta_2^{\sharp j} \kappa_\chi^{\sharp j} \chi_j(0)(1-\eta_2)}{\beta}, \frac{\kappa_\chi^{\sharp j} \chi_j(0)}{\beta_C}\right\}.
\end{aligned}$$

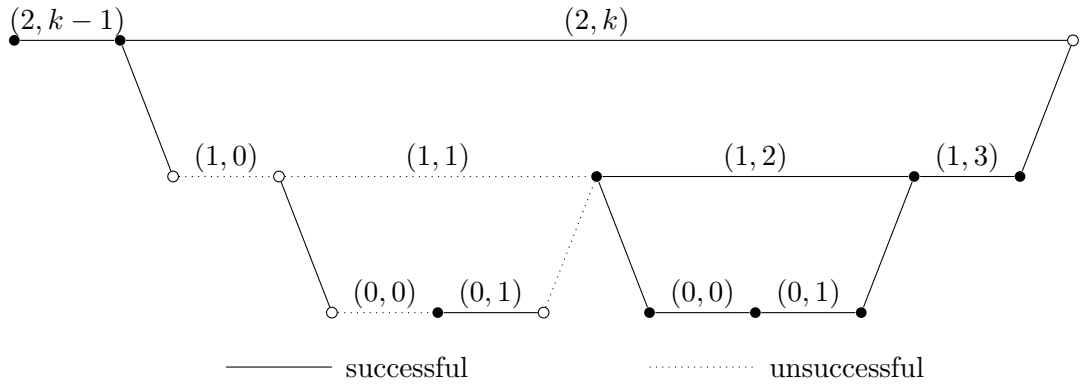

Figure 2.6.: Example iteration graph

This allows us to use the induction assumption at level j which yields that the first step $s_{j,0}$ is very successful and assertion (2.38) holds. Using (2.30) and $\sharp j \leq \sharp i - 1$, we can estimate the actual reduction on level j by

$$h_j(v_{j,0}) - h_j(v_{j,0} + s_{j,0}) \geq \eta_2^{\sharp j+1} \kappa_{\mathrm{mdc}}\, \kappa_\chi^{\sharp j} \chi_j(v_{j,0}) \Delta_{j,0} \geq \eta_2^{\sharp i} \kappa_{\mathrm{mdc}}\, \kappa_\chi^{\sharp i} \chi_i(v_{i,k}) \Delta_{i,k}. \tag{2.40}$$

Let us assume that we make $m \geq 0$ more steps on the j-th level and terminate afterwards. The algorithm is a descent method, which follows from the update rule in Step 4 of the algorithm. Therefore,

$$pred_{i,k} = h_j(v_{j,0}) - h_j(v_{j,m}) \geq h_j(v_{j,0}) - h_j(v_{j,0} + s_{j,0}).$$

From Lemma 2.6 and (2.39) we infer

$$\frac{pred_{i,k} - ared_{i,k}}{pred_{i,k}} \leq \frac{\beta \Delta_{i,k}}{\eta_2^{\sharp i} \kappa_{\mathrm{mdc}}\, \kappa_\chi^{\sharp i} \chi_i(x_i, v_{i,k})} \leq (1 - \eta_2)$$

and hence

$$\rho_{i,k} = \frac{ared_{i,k}}{pred_{i,k}} \geq \eta_2.$$

This shows that the step is very successful. Assumption (2.38) now follows immediately from (2.40). □

Remark 2.14 It is noteworthy that in the multilevel case the previous proof only uses the reduction of the first successful step on the coarser grid. This justifies the additional termination criteria after one successful step in Step 5 of the algorithm.

For the upcoming analysis we need to establish some additional notation. We say a multilevel iteration (i, k) *generates* another iteration (j, l) if (j, l) occurs in the recursion started and ended in iteration (i, k). Furthermore, let p be a function that returns the predecessor of a given iteration (j, l). This is either $(j, l - 1)$ if $l > 0$, or the multilevel iteration (i, k) in which level j was entered.

We are interested in every sub-step on lower levels of which the final multilevel step consists. Here, it is important that all steps generated by non successful multilevel iterations have no influence,

because the final step that they have contributed to is rejected. Thus we ignore these steps and denote by $\tilde{\mathcal{I}}(i,k)$ the chronological sequence of iterations that were generated by (i,k) without steps generated by non successful multilevel iterations. In case that (i,k) is a Taylor step, $\tilde{\mathcal{I}}(i,k)$ consists only of (i,k). An example with three levels is shown in Figure 2.6. Here, the sequence for iteration $(2,k)$ is

$$\tilde{\mathcal{I}}(2,k) = \big((2,k),(1,0),(1,1),(1,2),(0,0),(0,1),(1,3)\big).$$

The first two iterations $(0,0)$ and $(0,1)$ on level 0 are not included, because the multilevel step $(1,1)$ was not successful. Note that the numbering of the iterations is ambiguous since we normally enter a level more than once. In the following, it should be clear from the context, which iteration is meant.

We denote the first successful Taylor step of a sequence $\tilde{\mathcal{I}}(i,k)$ by $\alpha(i,k)$. The algorithm ensures that if (i,k) is successful, there is at least one successful Taylor iteration in every sequence $\tilde{\mathcal{I}}(i,k)$. This is because after entering a level j with $N(j) = \emptyset$ a successful Taylor step must be made before the algorithm is allowed to return. Furthermore, let $\mathcal{I}(i,k)$ the first part of $\tilde{\mathcal{I}}(i,k)$ until the step $\alpha(i,k)$. In the example iteration from Figure 2.6 we have $\alpha(2,k) = (0,0)$ and $\mathcal{I}(2,k) = \big((2,k),(1,0),(1,1),(1,2),(0,0)\big)$.

In the following we will omit the level index if we are on level r. We use a superscript to enumerate the tuples in the ordered sets $\mathcal{I}(i,k)$.

Remark 2.15 For $\mathcal{I}(i,k)$ holds:

$$\Delta_{\mathcal{I}(i,k)^{j+1}} \leq \Delta_{\mathcal{I}(i,k)^j} \leq \Delta_{i,k}, \quad j = 1,\dots,|\mathcal{I}(i,k)| - 1.$$

Furthermore, let $(j,l) \in \mathcal{I}(i,k)$ then

$$\chi_j(v_{j,l}) \geq \kappa_\chi^{\sharp i - \sharp j} \chi_i(v_{i,k}),$$

because $v_{j,l}$ is either $v_{i,k}$ if $j = i$, or $v_{j,l} = 0$ and condition (2.30) is satisfied.

Remark 2.16 If $(j,l) \in \mathcal{I}(i,k)$ is a successful multilevel iteration, then $\mathcal{I}(j,l) \subset \mathcal{I}(i,k)$.

The next lemma shows that if the stationarity measure is bounded below on a set of iterations, then the trust-region radius cannot become arbitrary small.

Lemma 2.8 *Let $\chi_i(v_{i,k}) \geq \varepsilon > 0$ for all iterations k on level i, then*

$$\begin{aligned} &\Delta_{i,k} \geq B_\Delta(\varepsilon) := \gamma_2 \min\left\{1, \kappa_{mdc} \frac{\kappa_\chi^{\sharp r} \eta_2^{\sharp r}(1-\eta_2)}{\beta}\varepsilon, \frac{\kappa_\chi^{\sharp r}}{\beta_C}\varepsilon\right\} && \text{if } i = r, \\ &\Delta_{i,k} \geq \min\{B_\Delta(\varepsilon), \varepsilon_i^\Delta \Delta_{i,0}\} && \text{if } i < r. \end{aligned} \tag{2.41a}$$

Moreover, for a multilevel step (i,k) we have for all $(j,l) \in \mathcal{I}(i,k)$:

$$\begin{aligned} &\Delta_{j,l} \geq B_\Delta(\varepsilon) && \text{if } i = r, \\ &\Delta_{j,l} \geq \min\{B_\Delta(\varepsilon), \varepsilon_i^\Delta \Delta_{i,0}\} && \text{if } i < r. \end{aligned} \tag{2.41b}$$

PROOF We first show (2.41a) for $i = r$. Suppose the statement of the lemma was false and the k-th iteration is the first one where

$$\Delta_k < B_\Delta(\varepsilon).$$

Then the preceding iteration must have been unsuccessful and it follows from the update rule (2.34) that

$$\Delta_{k-1} = \frac{\Delta_k}{\gamma_2} < \min\left\{1, \kappa_{\text{mdc}} \frac{\kappa_\chi^{\sharp r} \eta_2^{\sharp r}(1-\eta_2)}{\beta}\varepsilon, \frac{\kappa_\chi^{\sharp r}}{\beta_C}\varepsilon\right\}.$$

However, since $\varepsilon \leq \chi_r(v_{k-1})$, the fact that iteration $k-1$ is unsuccessful is a contradiction to Lemma 2.7 and therefore $\Delta_k \geq B_\Delta(\varepsilon)$.

We now turn to the case $i < r$, where we also assume that the assertion of the lemma is false and the k-th iteration is the first in which (2.41a) is violated. Since $\varepsilon_i^\Delta < 1$, the statement is obviously true for $k = 0$. If $k > 0$ and iteration $k-1$ is successful, it follows from (2.34) that

$$\Delta_{i,k} = \min\{c\Delta_{i,k-1}, \Delta_{i,0} - \|v_{i,k}\|_i\}$$

with $c = 1$ or $c = \gamma_1 > 1$. Since $\Delta_{i,k} < \Delta_{i,k-1}$, we conclude that

$$\Delta_{i,k} = \Delta_{i,0} - \|v_{i,k}\|_i.$$

From $\Delta_{i,k} < \varepsilon_i^\Delta \Delta_{i,0}$ follows $\varepsilon_i^\Delta \Delta_{i,0} > \Delta_{i,0} - \|v_{i,k} - v_{i,0}\|_i$. Hence, in iteration $k-1$ the termination criterion (2.35) was already satisfied contrary to the fact that there exists an iteration k. If, however, iteration $k-1$ is unsuccessful we get from (2.34), because of $v_{i,k-1} = v_{i,k}$ and $\gamma_2 < 1$, that $\Delta_{i,k} = \gamma_2 \Delta_{i,k-1}$. As in the case $i = r$, we can now derive a contradiction to Lemma 2.7. This completes the proof of (2.41a).

We also prove the last bound by contradiction. We assume that there exists a first iteration $(j,l) \in \mathcal{I}(i,k)$ where (2.41b) does not hold. From (2.41a) it follows that $(j,l) \neq (i,k)$. Furthermore $l > 0$, because otherwise the previous iteration $p(j,0)$ were the first one where the bound is violated (cf. Remark 2.9). From the definition of α it follows that $(j, l-1)$ is not a successful Taylor iteration. It also cannot be a successful multilevel iteration, since then there would have to be a successful Taylor step in $\mathcal{I}(j, l-1)$ and in this case $(j,l) \notin \mathcal{I}(i,k)$. Hence, it was unsuccessful. Using $\varepsilon \leq \chi_i(v_{i,k}) \leq \chi_j(v_{j,l-1})/\kappa_\chi^{\sharp i - \sharp j}$, which follows from Remark 2.15, one obtains

$$\begin{aligned}
\Delta_{j,l-1} &< \min\left\{1, \kappa_{\text{mdc}} \frac{\kappa_\chi^{\sharp r} \eta_2^{\sharp r}(1-\eta_2)}{\beta}\varepsilon, \frac{\kappa_\chi^{\sharp r}}{\beta_C}\varepsilon\right\} \\
&\leq \min\left\{1, \kappa_{\text{mdc}} \frac{\kappa_\chi^{\sharp r - \sharp i + \sharp j} \eta_2^{\sharp j}(1-\eta_2)}{\beta}\chi_j(v_{j,l-1}), \frac{\kappa_\chi^{\sharp r - \sharp i + \sharp j}}{\beta_C}\chi_j(v_{j,l-1})\right\} \\
&\leq \min\left\{1, \kappa_{\text{mdc}} \frac{\kappa_\chi^{\sharp j} \eta_2^{\sharp j}(1-\eta_2)}{\beta}\chi_j(v_{j,l-1}), \frac{\kappa_\chi^{\sharp j}}{\beta_C}\chi_j(v_{j,l-1})\right\}.
\end{aligned}$$

This is a contradiction, because again according to Lemma 2.7 iteration $(i, j-1)$ has to be successful. □

We next show that part of the descent that is obtained by the first successful Taylor step in a multilevel iteration carries over to the outgoing level.

Lemma 2.9 *Every successful iteration (i,k) leads to an actual reduction of*

$$h_i(v_{i,k}) - h_i(v_{i,k+1}) \geq \eta_1^{\sharp i - \sharp j + 1} \kappa_{mdc}\, \kappa_\chi^{\sharp i - \sharp j} \chi_i(v_{i,k}) \min\left[1, \frac{\kappa_\chi^{\sharp i - \sharp j}\chi_i(v_{i,k})}{\beta_C}, \Delta_{j,l}\right], \tag{2.42}$$

where $(j,l) = \alpha(i,k)$.

Proof Let us first suppose that (i,k) is a Taylor iteration. In this case, $(j,l) = (i,k)$ holds. By assumption, the step is successful and thus from the fraction of Cauchy decrease condition (2.29) it follows that

$$h_i(v_{i,k}) - h_i(v_{i,k+1}) \geq \eta_1 \kappa_{\mathrm{mdc}}\, \chi_i(v_{i,k}) \min\left[1, \frac{\chi_i(v_{i,k})}{\beta_C}, \Delta_{i,k}\right].$$

Since $\kappa_\chi < 1$ and $\eta_1 < 1$, (2.42) is proven in this case.

Now let $j \prec i$. From the definition of α, it follows that the step (j,l) is the first successful one on level j and a Taylor step. Due to this, because of $v_{j,l} = v_{j,0}$ and (2.29), we obtain the actual reduction

$$h_j(v_{j,0}) - h_j(v_{j,l+1}) \geq \eta_1 \kappa_{\mathrm{mdc}}\, \chi_j(v_{j,0}) \min\left[1, \frac{\chi_j(v_{j,0})}{\beta_C}, \Delta_{j,l}\right].$$

The algorithm is a descent method and therefore the reduction achieved by the final step on level j, $s_j^* = v_{j,*} - v_{j,0}$, is also greater than or equal to the right hand side of the last inequality. According to the definition of $\mathcal{I}(i,k)$, the prolongation of the step s_j^* is successful. Let $(\bar{j},\bar{l}) = p(j,0)$. For the iteration $(\bar{j},\bar{l})$ to be valid, (2.30) must have been satisfied. This yields

$$\begin{aligned} h_{\bar{j}}(v_{\bar{j},\bar{l}}) - h_{\bar{j}}(v_{\bar{j},\bar{l}} + P_{\bar{j}}^{\bar{j}} s_{\bar{j}}^*) &= h_{\bar{j}}(v_{\bar{j},\bar{l}}) - h_{\bar{j}}(v_{\bar{j},\bar{l}+1}) \\ &\geq \eta_1^2 \kappa_{\mathrm{mdc}}\, \kappa_\chi \chi_{\bar{j}}(v_{\bar{j},\bar{l}}) \min\left[1, \frac{\kappa_\chi \chi_{\bar{j}}(v_{\bar{j},\bar{l}})}{\beta_C}, \Delta_{j,l}\right]. \end{aligned}$$

If $\bar{j} = i$, then $\bar{l} = k$ and the proof were completed. Otherwise if $\bar{j} \prec i$ we know from the definition of α that $\bar{l}$ is the first successful iteration on $\bar{j}$ and therefore $v_{\bar{j},\bar{l}} = v_{\bar{j},0}$. The rest of the proof follows straightforwardly by applying the above arguments inductively. For every level in the sequence between i and j we get the additional factors κ_χ and η_1 which explains the factor $\kappa_\chi^{\sharp i - \sharp j}$ and $\eta_1^{\sharp i - \sharp j + 1}$ in (2.42). □

Up to now, we have always assumed that it is possible to generate multilevel steps, which means that if we make a multilevel step, at least one termination criterion of the algorithm is satisfied after a finite number of iterations on the lower levels. The next lemma shows that this is indeed guaranteed.

Lemma 2.10 *Let all lower-level models h_i be bounded below on all balls $\{v_i \in C_i \,|\, \|v_i\|_i \leq \Delta\}$ with $0 \leq \Delta < \infty$. Then every multilevel step (i,k) is well defined, i.e., always generates only a finite number of iterations on the lower levels.*

PROOF We first show that we only make a finite number of iterations if we enter a level i with $N(i) = \emptyset$, i.e., a level where every step is a Taylor step. Suppose the assertion is false, then for every iteration (j, l) the termination criterion is not satisfied and therefore

$$\chi_j(v_{j,l}) > \varepsilon_j^\chi \quad \text{and} \quad \|v_{j,l} - v_{j,0}\|_j < (1 - \varepsilon_j^\Delta)\Delta_{j,0} \quad \text{for all iterations } l.$$

From Lemma 2.8 follows $\Delta_{j,l} \geq \min\{B_\Delta(\varepsilon_j^\chi), \varepsilon_j^\Delta \Delta_{j,0}\} =: C$ and consequently we make infinitely many successful steps. Every successful step satisfies the fraction of Cauchy decrease condition (2.29), so we can estimate the actual reduction by

$$ared_{j,l} = h_j(v_{j,l}) - h_j(v_{j,l+1}) \geq \eta_1 \, \kappa_{\text{mdc}} \, \varepsilon_j^\chi \min\left[1, \varepsilon_j^\chi/\beta_C, C\right].$$

Let $\theta(l)$ be the number of successful steps till the $l - th$ iteration, then we get

$$\begin{aligned} h_j(v_{j,0}) - h_j(v_{j,l}) &= \sum_{\nu=0}^{l-1} (h_j(v_{j,\nu}) - h_j(v_{j,\nu+1})) \\ &\geq \theta(k) \, \eta_1 \, \kappa_{\text{mdc}} \, \varepsilon_j^\chi \min\left[1, \varepsilon_j^\chi/\beta_C, C\right] \to \infty \text{ for } k \to \infty. \end{aligned}$$

Because all iterates lie in the set $\{v_j \in C_j \,|\, \|v_j - v_{j,0}\|_j \leq \Delta_{j,0}\}$, which is a subset of the ball $\{v_j \in C_j \,|\, \|v_j\|_j \leq \Delta_{j,0} + \|v_{j,0}\|_j\}$, this is a contradiction to the boundedness from below of h_j on balls.

Now we suppose that the assumption holds for all multilevel iterations on level j that was entered in iteration (i, k). Again, we assume that the termination criteria in Step 5 of Algorithm 1 are never satisfied. As in the case $N(j) = \emptyset$, it follows from Lemma 2.8 that all trust-region radii $\Delta_{j,l}$ are bounded below by a constant C and therefore we make infinite many successful steps. From the induction assumption we already now that every multilevel iteration is finished after a finite amount of time. So it suffices to show that we only make a finite number of iterations on level j. For a successful iteration (j, l), it follows from Lemma (2.9) that

$$h_j(v_{j,l}) - h_j(v_{j,l+1}) \geq \eta_1^{\sharp j - \sharp \bar{j} + 1} \kappa_{\text{mdc}} \, \kappa_\chi^{\sharp j - \sharp \bar{j}} \chi_j(v_{j,l}) \min\left[1, \frac{\kappa_\chi^{\sharp j - \sharp \bar{j}} \chi_j(v_{j,l})}{\beta_C}, \Delta_{\bar{j},\bar{l}}\right]$$

with $(\bar{j}, \bar{l}) = \alpha(j, l)$. According to the second assertion of Lemma 2.8, $\Delta_{\bar{j},\bar{l}} \geq C$ and thus with $\chi_j(v_{j,l}) \geq \varepsilon_j^\chi$

$$h_j(v_{j,l}) - h_j(v_{j,l+1}) \geq C'$$

with a constant C' that does not depend on k. By the same argument as in the case $N(i) = \emptyset$, we can derive a contradiction to the boundedness of h_j on balls and the lemma is proven. □

Remark 2.17 The previous lemma is obviously satisfied without any further assumptions if we add an additional termination condition in Step 5 of the algorithm: Return when $i < r$ and the number of successful steps $\theta(k)$ on this level is greater or equal a fixed constant $k_{\max} \in \mathbb{N}$.

We will now analyse the convergence behavior of the algorithm on the finest level. To this end, we assume that $\varepsilon_r^\chi = 0$ and we show that the sequence $(\chi_r(v_{r,k}))_{k \in \mathbb{N}}$ generated by Algorithm 2.1 converges to zero. We first prove that, provided there are only finitely many successful iterations, the last successful iteration belongs to a stationary point.

Lemma 2.11 *Let $(v_k)_{k\in\mathbb{N}}$ be a sequence generated by Algorithm 2.1. Suppose that $\varepsilon_r^\chi = 0$ and that there are only finitely many successful iterations on the finest grid. Then $v_k = v_*$ for sufficiently large k and $\chi_r(v_*) = 0$.*

Proof Assume that the algorithm generates infinitely many iterations. From the assumptions follows the existence of a last successful iteration on the finest grid, which we denote by $(r, *)$. Since all remaining iterations are unsuccessful, $\gamma_2 < 1$ implies $\Delta_{r,k} \to 0,\ k \to \infty$ and $v_{r,k} = v_{r,*}$ for $k > *$. If $\chi_r(v_{r,*}) > 0$, it follows from Lemma 2.7 that there exists a successful iteration (r, k) with $k > *$, which is contrary to the assumption. Hence $\chi_r(v_{r,*}) = 0$. □

If we make infinitely many successful steps, the next result states that there is at least one subsequence that converges to a stationary point.

Theorem 2.1 *Let f_r be bounded below on C_r and let $(v_k)_\mathbb{N}$ be a sequence generated by Algorithm 2.1. Furthermore, let $\varepsilon_r^\chi = 0$. If the algorithm does not terminate after a finite number of iterations, then*

$$\liminf_{k\to\infty} \chi_r(v_{r,k}) = 0. \tag{2.43}$$

Proof Lemma 2.11 implies that the algorithm generates infinitely many successful steps. Suppose that the assumption does not hold. Then there exists an $\varepsilon > 0$ such that

$$\chi_r(v_{r,k}) \geq \varepsilon \quad \text{for all } k.$$

Hence, Lemma 2.8 gives a lower bound on the trust-region radii Δ_k. Similar to the second part of the proof of Lemma 2.10 one shows that

$$\lim_{k\to\infty} (f_r(x_r + v_{r,0}) - f_r(x_r + v_{r,k})) \geq C' \lim_{k\to\infty} \theta(k) = \infty$$

where $\theta(k)$ denotes the number of successful steps until the kth iteration. Because $v_{r,k} \in C_r$ for all k, this is a contradiction to the boundedness of f_r on C_r. □

Lemma 2.12 *The descent of a successful step k on level r satisfies*

$$f_r(x_r + v_{r,k}) - f_r(x_r + v_{r,k+1}) \geq \eta_1^{\sharp r+1} \kappa_{mdc}\, \kappa_\chi^{\sharp r} \chi_r(v_{r,k}) \min\left[\Delta_{r,k}, B_\Delta(\chi_r(v_{r,k}))\right],$$

where B_Δ is defined as in (2.41a).

Proof On level r, the model $h_r(v_{r,k})$ is equal to the function $f_r(x_r + v_{r,k})$. Because iteration (r, k) is successful, we use estimate (2.42) from Lemma 2.9 with $(j, l) = \alpha(r, k)$ to obtain

$$\begin{aligned} f_r(x_r + v_{r,k}) - f_r(x_r + v_{r,k+1}) &\geq \eta_1^{\sharp r-\sharp j+1} \kappa_{\mathrm{mdc}}\, \kappa_\chi^{\sharp r-\sharp j} \chi_r(v_{r,k}) \min\left[1, \Delta_{j,l}, \frac{\kappa_\chi^{\sharp r-\sharp l} \chi_r(v_{r,k})}{\beta_C}\right] \\ &\geq \eta_1^{\sharp r+1} \kappa_{\mathrm{mdc}}\, \kappa_\chi^{\sharp r} \chi_r(v_{r,k}) \min\left[\Delta_{j,l}, B_\Delta(\chi_r(v_{r,k})), \Delta_{r,k}\right], \end{aligned} \tag{2.44}$$

where the second inequality follows from $B_\Delta(\chi_r(v_{r,k})) \leq 1$, $B_\Delta(\chi_r(v_{r,k})) \leq \frac{\kappa_\chi^{\sharp r} \chi_r(v_{r,k})}{\beta_C}$ and $\Delta_{j,l} \leq \Delta_{r,k}$. Without loss of generality, we can demand that either $j = r$ or $l > 0$, because otherwise since $\Delta_{j,0} = \Delta_{p(j,0)}$, we can replace (j, l) by $p(j, 0)$ in (2.44) as long as $j \prec r$ and $l = 0$.

If $\Delta_{j,l} \geq B_\Delta(\chi_r(v_{r,k}))$ or $j = r$, the assertion is true. Let us now suppose $\Delta_{j,l} < B_\Delta(\chi_r(v_{r,k}))$ and $l > 0$. The definition of the function α implies that iteration $(j, l-1)$ was unsuccessful and hence

$$\Delta_{j,l-1} < \frac{B_\Delta(\chi_r(v_{r,k}))}{\gamma_2}.$$

After inserting the definition of B_Δ and using $\chi_j(v_{j,l-1}) = \chi_j(v_{j,l}) \geq \kappa_\chi^{\sharp r - \sharp j} \chi_r(v_{r,k})$, we obtain

$$\Delta_{j,l-1} < \min\left\{1, \kappa_{\mathrm{mdc}} \frac{\kappa_\chi^{\sharp j} \eta_2^{\sharp j}(1-\eta_2)}{\beta} \chi_j(v_{j,0}), \frac{\kappa_\chi^{\sharp j} \chi_j(v_{j,0})}{\beta_C}\right\}.$$

Therefore, the unsuccessfulness of step $v_{j,l-1}$ is a contradiction to Lemma 2.7 and it follows that $\Delta_{j,l} \geq B_\Delta(\chi_r(v_{r,k}))$. □

Now we can prove the global convergence of the algorithm under the additional assumption that the stationarity measure χ_r is uniformly continuous on a suitable subset of C_r.

Theorem 2.2 *Let f_r be bounded below on C_r and let χ_r be uniformly continuous on a set $\mathcal{S} \subset C_r$ that contains the sequence of iterates $(v_{r,k})_{k\in\mathbb{N}}$. Then*

$$\lim_{k\to\infty} \chi_r(v_{r,k}) = 0. \tag{2.45}$$

Proof We denote by S the set of successful iterations on level r.

Let us assume that (2.45) is not true. Then there exists $\varepsilon > 0$ such that $\chi_r(v_{r,k}) \geq 2\varepsilon$ for infinitely many $k \in S$. Since (2.43) holds, we thus find increasing sequences $(j_i')_{i\geq 0} \subset S$ and $(k_i')_{i\geq 0} \subset S$ with $j_i' < k_i' < j_{i+1}'$ and

$$\chi_r(v_{r,j_i'}) \geq 2\varepsilon \quad \chi_r(v_{r,k}) > \varepsilon \;\forall\, k \in S \text{ with } j_i' < k < k_i', \quad \chi_r(v_{r,k_i'}) \leq \varepsilon.$$

Setting $S' = \bigcup_{i=0}^\infty S_i'$ with $S_i' = \{k \in S;\, j_i' \leq k < k_i'\}$, we have

$$\liminf_{S'\ni k\to\infty} \chi_r(v_{r,k}) \geq \varepsilon.$$

Using Lemma 2.12 we deduce for $k \in S'$ that

$$f_r(x_r + v_{r,k}) - f_r(x_r + v_{r,k+1}) \geq \eta_1^{\sharp r+1} \kappa_{\mathrm{mdc}}\, \kappa_\chi^{\sharp r} \varepsilon \min[\Delta_{r,k}, B_\Delta(\varepsilon)]. \tag{2.46}$$

The sequence $\{f_r(x_r+v_{r,k})\}_k$ is monotonically decreasing and bounded below, hence it is convergent and the left-hand side of (2.46) must tend to zero when k tends to infinity. This gives

$$\lim_{S'\ni k\to\infty} \Delta_k = 0.$$

As a consequence, the first term dominates in the minimum of (2.46) and we obtain that, for $k \in S'$ sufficiently large,

$$\Delta_k \leq \frac{1}{\eta_1^{\sharp r+1} \kappa_{\mathrm{mdc}}\, \kappa_\chi^{\sharp r} \varepsilon} [f_r(x_r + v_{r,k}) - f_r(x_r + v_{r,k+1})].$$

We then deduce from this bound that, for i sufficiently large,

$$\begin{aligned}\|v_{r,j'_i} - v_{r,k'_i}\|_r &\leq \sum_{j \in S',\, j=j'_i}^{k'_i} \|v_{r,j} - v_{r,j+1}\|_r \leq \sum_{j \in S',\, j=j'_i}^{k'_i} C_{\mathcal{P}} \Delta_{r,j} \\ &\leq \frac{C_{\mathcal{P}}}{\eta_1^{\sharp r+1} \kappa_{\mathrm{mdc}}\, \kappa_\chi^{\sharp r}\, \varepsilon} [f_r(x_r + v_{r,j'_i}) - f_r(x_r + v_{r,k'_1})].\end{aligned}$$

The right hand side of this inequality must converge to zero, and therefore $\|v_{r,j'_i} - v_{r,k'_i}\|_r$ tends to zero as i tends to infinity. By uniform continuity of χ_r, we thus deduce that $\chi_r(v_{r,j'_i}) - \chi_r(v_{r,k'_i})$ tends to zero. However this is impossible, because of the definition of (j'_i) and (k'_i), which imply that $\chi_r(v_{r,j'_i}) - \chi_r(v_{r,k'_i}) \geq \varepsilon$. □

3. Unconstrained problems

In this chapter we consider unconstrained problems, i.e., where $C_i = \mathcal{V}_i$ holds, in a typical multilevel setting. Since $h_i'(v_{i,k}) \in \mathcal{V}_i^*$, the natural stationarity measure is the (dual)-norm of the derivative, i.e.,

$$\chi_i(v_{i,k}) = \|h_i'(v_{i,k})\|_{\mathcal{V}_i^*}. \tag{3.1}$$

We assume that the spaces $\mathcal{V}_i$, $i = 1, \dots, r$, are finite dimensional and subsets of a suitable chosen Hilbert space $\mathcal{U}$. In this setting we will first show an important norm equivalence if the smoothness property (2.30) does not hold. Further, we analyze different possibilities how to implement the Taylor step computation in the trust-region algorithm. In the case of convex trust-region subproblems, we show that classical smoothing algorithms, like Gauß-Seidel or Jacobi smoothers, can be used to calculate an approximate solution. Our main result is that provided the smoothness assumption (2.30) is not satisfied, a typical smoothing step achieves a descent satisfying the fraction of Cauchy decrease condition (2.29) where the constant κ_{mdc} is independent of the level i.

Throughout this chapter we will use a generic constant C which neither depends on the level i nor the number of levels r. C may assume different values in the inequalities and is assumed to be large enough, such that the inequality is satisfied. In general we call a quantity *level-independent* if it does not depend on the level i and also does not deteriorate for $r \to \infty$.

3.1. The variational setting

Let $\mathcal{U}$ be a Hilbert space with an inner product $(\cdot, \cdot)$ and associated norm $\|\cdot\| = \sqrt{(\cdot, \cdot)}$. Furthermore, let $\mathcal{V} \hookrightarrow \mathcal{U}$ be a dense and continuously embedded Hilbert subspace with inner product $(\cdot, \cdot)_{\mathcal{V}}$. Then $\mathcal{V} \subset \mathcal{U} \subset \mathcal{V}^*$ forms a *Gelfand triple* (cf. Section 2.1.2).

We assume that we have a nested sequence of finite dimensional subspaces $\mathcal{V}_1 \subset \mathcal{V}_2 \subset \ldots \subset \mathcal{V}_r \subset \mathcal{V}$ with dimensions $n_1, \dots, n_r$ and norms $\|\cdot\|_{\mathcal{V}_i} := \|\cdot\|_{\mathcal{V}}$. We suppose a multilevel hierarchy as in Remark 2.1. Furthermore, let $\{\phi_i^j\}_{j=1}^{n_i}$ be a basis of $\mathcal{V}_i$ for every $i = 1, \dots, r$. Every element $v_i \in \mathcal{V}_i$ can be represented by $v_i = \sum_{j=1}^{n_i} \tilde{v}_i^j \phi_i^j$ where $\tilde{v}_i \in \mathbb{R}^{n_i}$ denotes the associated coefficient vector. As in Example 2.1, we suppose that the identity between $\mathcal{V}_i$ and $\mathcal{V}_{i+1}$ is used as prolongation P_i^{i+1}. We will often regard an element of $\mathcal{U}$ as element of its dual space by means of the embedding $\iota_{\mathcal{U}} \colon \mathcal{U} \to \mathcal{U}^*$, $v \mapsto (v, \cdot)$.

In the following, we need the $\mathcal{U}$-orthogonal projection onto the space $\mathcal{V}_i$, which we denote by $Q_i \colon \mathcal{U} \to \mathcal{V}_i$. According to Theorem A.2, it satisfies the relation

$$(Q_i u, v_i) = (u, v_i) \quad \text{for all } v_i \in \mathcal{V}_i \text{ and } u \in \mathcal{U}. \tag{3.2}$$

In this setting, there exist representations $g_{i,k} \in \mathcal{V}_i$ and $H_{i,k} \colon \mathcal{V}_i \to \mathcal{V}_i$ such that the quadratic function $q_{i,k}$ of the trust-region subproblem (2.32) can be written as

$$q_{i,k}(s_{i,k}) = (s_{i,k},\, g_{i,k}) + \frac{1}{2}(s_{i,k},\, H_{i,k}s_{i,k}).$$

This is shown by the following lemma:

Lemma 3.1 *Let $g \in \mathcal{V}_i^*$ and $\mathcal{V}_i$ a finite dimensional subspace of a Hilbert space $\mathcal{U}$. Then there exists an element $g_i \in \mathcal{V}_i$ such that*

$$(v_i, g_i) = \langle g, v_i \rangle \quad \forall\, v_i \in \mathcal{V}_i. \tag{3.3}$$

Proof Since $\mathcal{V}_i$ is finite dimensional, $\mathcal{V}_i$ equipped with the inner product $(\cdot,\cdot)$ forms a Hilbert space. From the Riesz representation theorem follows the existence of an element g_i that satisfies (3.3). □

Remark 3.1 The choice of g_i does not seem to be natural when $\mathcal{V}$ is a Hilbert space. Instead one would like to use the representation with regard to $(\cdot,\cdot)_{\mathcal{V}}$. The main difficulty lies in the fact that the calculation of this representation is often expensive whereas the one of Lemma 3.1 comes for free in many applications. See Chapter 5 for details. This is a major difference to *Sobolev gradient methods* where, in case that $\mathcal{V}$ is a Sobolev space, a gradient representation with regard to $(\cdot,\cdot)_{\mathcal{V}}$ is used, cf., e.g., [Neu97].

Let $g_i = \nabla_{\mathcal{U}} h_i(v_{i,k})$ be the representative of $h_i'(v_{i,k})$ according to Lemma 3.1. The representation of the adjoint of the prolongation operator $P_{i-1}^i \colon \mathcal{V}_{i-1} \to \mathcal{V}_i$ is given by the $\mathcal{U}$-orthogonal projection Q_{i-1} since

$$\langle (P_{i-1}^i)^* h_i'(v_{i,k}), v_{i-1} \rangle = \langle h_i'(v_{i,k}), v_{i-1} \rangle = (g_i, v_{i-1}) = (Q_{i-1}g_i, v_{i-1}) = \langle \iota_{\mathcal{U}}(Q_{i-1}g_i), v_{i-1} \rangle.$$

Furthermore, due to the choice of the stationarity measure, it follows directly from Definition 2.1 of the lower-level models that

$$\begin{aligned}\chi_{i-1}(0) = \|h_{i-1}'(0)\|_{\mathcal{V}_{i-1}^*} &= \sup_{v_{i-1} \in \mathcal{V}_{i-1}} \frac{\langle h_{i-1}'(0), v_{i-1} \rangle}{\|v_{i-1}\|_{\mathcal{V}_{i-1}}} = \sup_{v_{i-1} \in \mathcal{V}_{i-1}} \frac{\langle (P_{i-1}^i)^* h_i'(v_{i,k}), v_{i-1} \rangle}{\|v_{i-1}\|_{\mathcal{V}_{i-1}}} \\ &= \sup_{v_{i-1} \in \mathcal{V}_{i-1}} \frac{(Q_{i-1}g_i, v_{i-1})}{\|v_{i-1}\|_{\mathcal{V}_{i-1}}} = \|\iota_{\mathcal{U}}(Q_{i-1}g_i)\|_{\mathcal{V}_{i-1}^*}.\end{aligned}$$

As a tool for our analysis, we define for $i = 1, \ldots, r$ the linear operators $V_i \colon \mathcal{V}_i \to \mathcal{V}_i$ by

$$(V_i v_i, w_i) = (v_i, w_i)_{\mathcal{V}} \quad \text{for all } v_i, w_i \in \mathcal{V}_i. \tag{3.4}$$

Remark 3.2 The operators V_i satisfy $V_i = Q_i V_r$ because

$$(V_i v_i, w_i) = (v_i, w_i)_{\mathcal{V}} = (V_r v_i, w_i) = (Q_i V_r v_i, w_i).$$

From the definition it follows that the operators V_i are symmetric and positive definite, i.e.,

$$\begin{aligned}(V_i v_i, w_i) &= (v_i, V_i w_i) \quad \text{for all } v_i, w_i \in \mathcal{V}_i,\\ (v_i, V_i v_i) &= \|v_i\|_{\mathcal{V}}^2 > 0 \quad \text{for all } 0 \neq v_i \in \mathcal{V}_i.\end{aligned}$$

Therefore, the powers V_i^s, $s \in \mathbb{R}$, are well-defined and we can define a scale of norms by

$$|||v|||_{i,s} := \sqrt{(V_i^s v, v)}. \tag{3.5}$$

Directly from the definition it follows $|||v_i|||_{i,0} = \|v_i\|$ and $|||v_i|||_{i,1} = \|v_i\|_{\mathcal{V}}$ for $v_i \in \mathcal{V}_i$.

The next lemma shows that the dual norms of $\|\cdot\|_{\mathcal{V}_i}$ and $\|\cdot\|_{\mathcal{V}}$ are equivalent on the space $\{\iota_{\mathcal{U}}(v_i) \,|\, v_i \in \mathcal{V}_i\}$ for suitable spaces $\mathcal{U}$:

Lemma 3.2 *1. For all $i = 1, \dots, r$ and $g_i \in \mathcal{V}_i$ we have $|||g_i|||_{i,-1} = \|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*}$.*

2. Let the projection Q_i be stable in $\mathcal{V}$, i.e., there exists a level-independent constant $C_Q \geq 0$ such that

$$\|Q_i v\|_{\mathcal{V}} \leq C_Q \|v\|_{\mathcal{V}} \quad \text{for all } v \in \mathcal{V}. \tag{3.6}$$

Then the norms $\|\cdot\|_{\mathcal{V}_i^*}$ and $\|\cdot\|_{\mathcal{V}^*}$ are equivalent on $\mathcal{V}_i$, more precisely

$$\|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*} \leq \|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}^*} \leq C_Q \|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*} \quad \text{for all } g_i \in \mathcal{V}_i. \tag{3.7}$$

PROOF 1. Let $g_i \in \mathcal{V}_i$. We first remark that $\|v_i\|_{\mathcal{V}} = \|V_i^{1/2} v_i\|$ for $v_i \in \mathcal{V}_i$. From the definition of the dual norm we infer

$$\|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*} = \sup_{v_i \in \mathcal{V}_i} \frac{(g_i, v_i)}{\|v_i\|_{\mathcal{V}}} = \sup_{v_i \in \mathcal{V}_i} \frac{(V_i^{-1/2} g_i, V_i^{1/2} v_i)}{\|V_i^{1/2} v_i\|}.$$

Since $v_i \mapsto V_i^{1/2} v_i$ is surjective, we have

$$\|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*} = \sup_{w_i \in \mathcal{V}_i} \frac{(V_i^{-1/2} g_i, w_i)}{\|w_i\|} = \|V_i^{-1/2} g_i\| = |||g_i|||_{i,-1}.$$

2. Using the definition of the dual norm and that $\mathcal{V}_i \subset \mathcal{V}$ we obtain

$$\|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*} = \sup_{v_i \in \mathcal{V}_i} \frac{(g_i, v_i)}{\|v_i\|_{\mathcal{V}_i}} \leq \sup_{v \in \mathcal{V}} \frac{(g_i, v)}{\|v\|_{\mathcal{V}}} = \|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}^*}.$$

To verify the second inequality, we use (3.2) and the stability of Q_i:

$$\|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}^*} = \sup_{v \in \mathcal{V}} \frac{(g_i, v)}{\|v\|_{\mathcal{V}}} \leq C_Q \sup_{v \in \mathcal{V}} \frac{(g_i, Q_i v)}{\|Q_i v\|_{\mathcal{V}}} = C_Q \sup_{v \in \mathcal{V}_i} \frac{(g_i, v_i)}{\|v_i\|_{\mathcal{V}}} = C_Q \|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*}.$$

□

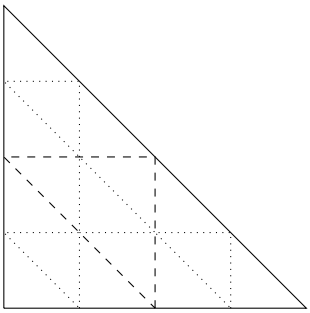

Figure 3.1.: Twice uniform refinement of a triangle

We do not assume that the norms on $\mathcal{U}$ and $\mathcal{V}$ are equivalent. Thus, on the finite dimensional spaces $\mathcal{V}_i$ the equivalence constants of these norms are in general level-dependent. We demand that the constants do not grow too fast, i.e., there exists a constant τ, independent of i, such that

$$\frac{\lambda_i^{\max}}{\lambda_{i-1}^{\max}} \leq \tau, \quad \text{for all } i = 1, \ldots, r, \tag{3.8}$$

where

$$\lambda_j^{\max} := \sup_{v_j \in \mathcal{V}_j} \frac{\|v_j\|_{\mathcal{V}}^2}{\|v_j\|^2}. \tag{3.9}$$

Without loss of generality we assume $\lambda_j^{\max} \geq 1$.

The following example describes a typical setting which we will often consider throughout this thesis:

Example 3.1 Let $\Omega \subset \mathbb{R}^d$ be a bounded polygonal domain, $\mathcal{V} = H_0^1(\Omega)$ and $\mathcal{U} = L^2(\Omega)$. It is well known that $H_0^1(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$ forms a Gelfand triple. As in Example 2.1 let $\mathcal{T}_1 \subset \{T_{h_1}\}$ be a conforming triangulation of $\overline{\Omega}$ with simplices of diameter less than or equal to h_1. We assume that the family of triangulations $\{T_{h_1}\}$ is quasi-uniform, i.e., there are constants $\sigma_1, \sigma_2 > 0$ such that

$$\max_{t \in \mathcal{T}_1} \frac{h_t}{\rho_t} \leq \sigma_1, \quad \frac{\max_{t \in \mathcal{T}_1} h_t}{\min_{t \in \mathcal{T}_1} h_t} \leq \sigma_2 \quad \forall h_1 \geq 0, \tag{3.10}$$

where h_t denotes the diameter of t and ρ_t the diameter of the largest ball contained in t. Let $\mathcal{N}_1$ be the set of nodes of $\mathcal{T}_1$ that are not on the boundary $\partial\Omega$. We create a sequence $\mathcal{T}_1, \ldots, \mathcal{T}_r$ with corresponding node sets $\mathcal{N}_1, \ldots, \mathcal{N}_r$ obtained from $\mathcal{T}_1$ by regular subdivision (cf. Figure 3.1). Therefore, with $h_j = \max_{t \in \mathcal{T}_j} \operatorname{diam}(t)$, we have the following relation between the mesh sizes:

$$h_1 = 2^{j-1} h_j.$$

On each triangulation we define a finite element space $\mathcal{V}_i$ that consists of continuous functions which are linear on each triangle $t \in \mathcal{T}_i$ and vanish on $\partial\Omega$. Since the triangulations are nested, we have

$$\mathcal{V}_1 \subset \mathcal{V}_1 \subset \ldots \subset \mathcal{V}_r \subset H_0^1(\Omega).$$

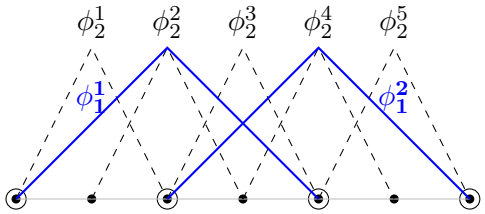

Figure 3.2.: The nodal basis functions for two consecutive levels in one dimension

For each node $x_i^j \in \mathcal{N}_i$, there exists a unique function $\phi_i^j \in \mathcal{V}_i$ satisfying

$$\phi_i^j(x_i^k) = \delta_{jk} \quad \text{for all } x_i^k \in \mathcal{N}_i. \tag{3.11}$$

The set $\{\phi_i^j\}_{j=1}^{n_i} \subset W_0^{1,\infty}(\Omega)$ forms a basis of $\mathcal{V}_i$. A basis satisfying (3.11) for $j = 1, \dots, n_i$ will in the following be referred to as *nodal basis*.

The largest eigenvalue of V_i can be estimated by

$$\lambda_i^{\max} = \sup_{v_i \in \mathcal{V}_i} \frac{\|v_i\|_{H^1(\Omega)}^2}{\|v_i\|_{L^2(\Omega)}^2} \leq C h_i^{-2}, \tag{3.12}$$

which follows directly from an inverse inequality (see for instance [Cia78, Thm. 3.2.6]). This upper bound cannot be improved, which can be seen by setting $v_i = \phi_i^j$ in the above fraction. Therefore, assumption (3.8) is fulfilled in this setting. The H_0^1-stability of the L^2-orthogonal projector, necessary for (3.6) to hold, is a well known fact. A rigorous proof can be found for instance in [BX91, Thm. 3.4].

3.2. Level-independent Cauchy decrease

In this section we show that under certain assumptions the fraction of Cauchy decrease condition (2.29) is satisfied by a very simple and cheap smoothing step. We are in particular interested in a decrease that is independent of the number of levels and the mesh size h of the discretizations. This was not examined in other multilevel optimization works, e.g., [GST08, GMTWM08, WG09, Nas00] where level dependent factors like the Euclidean norm of the stiffness matrices or the dimensions of the finite element spaces appear in estimates.

We will first analyse how the violation of the smoothness property allows us to derive an estimate for the dual norms. This is done in two cases.

3.2.1. The regular case

We will first assume that a strong regularity assumption is satisfied. We need the $\mathcal{V}_i$-orthogonal projection, which we denote by P_i. We use this notation, although it is similar to the prolongation

operators, since it is quite common in the literature and is only needed in this section. The operator $P_i\colon \mathcal{V} \to \mathcal{V}_i$ is defined by the relation

$$(P_i u, v_i)_{\mathcal{V}} = (u, v_i)_{\mathcal{V}} \quad \text{for all } v_i \in \mathcal{V}_i.$$

We assume that it also satisfies the relation

$$\|e_i - P_{i-1}e_i\|^2 \leq C(\lambda_{i-1}^{\max})^{-1}\|e_i - P_{i-1}e_i\|_{\mathcal{V}}^2 \quad \text{for all } e_i \in \mathcal{V}_i \text{ and } i = 2, \dots, r. \tag{3.13}$$

We will later discuss when this assumption holds in the setting of Example 3.1.

For the following lemma we use the identity $Q_{i-1}V_i = V_{i-1}P_{i-1}|_{\mathcal{V}_i}$, which can be shown easily: Let $e_i \in \mathcal{V}_i$, then for all $v_{i-1} \in \mathcal{V}_{i-1}$ we have

$$(V_{i-1}P_{i-1}e_i, v_{i-1}) = (P_{i-1}e_i, v_{i-1})_{\mathcal{V}} = (e_i, v_{i-1})_{\mathcal{V}} = (V_i e_i, v_{i-1}) = (Q_{i-1}V_i e_i, v_{i-1}).$$

Lemma 3.3 *Let* (3.13) *be satisfied and let $g_i \in \mathcal{V}_i$ be not smooth, i.e., it holds:*

$$\|\iota_{\mathcal{U}}(Q_{i-1}g_i)\|_{\mathcal{V}_{i-1}^*} < \kappa_\chi \|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*}. \tag{3.14}$$

Then there exists a level-independent constant C such that the following estimate is satisfied:

$$\|g_i\|^2 \geq C^{-1}\tau^{-1}(1 - \kappa_\chi^2)\lambda_i^{\max}\|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*}^2 \tag{3.15}$$

Proof Set $e_i = V_i^{-1}g_i$. The element $e_i - P_{i-1}e_i$ is $\mathcal{V}$-orthogonal on $\mathcal{V}_{i-1}$, hence

$$\|e_i - P_{i-1}e_i\|_{\mathcal{V}}^2 = (e_i - P_{i-1}e_i, e_i - P_{i-1}e_i)_{\mathcal{V}} = (e_i - P_{i-1}e_i, V_i e_i) \leq \|e_i - P_{i-1}e_i\|\|V_i e_i\|.$$

Inserting the approximation property (3.13) yields

$$\|e_i - P_{i-1}e_i\|_{\mathcal{V}}^2 \leq C^{1/2}(\lambda_{i-1}^{\max})^{-1/2}\|e_i - P_{i-1}e_i\|_{\mathcal{V}}\|g_i\|.$$

After dividing by $\|e_i - P_{i-1}e_i\|_{\mathcal{V}}$ and using (3.8), we obtain

$$\|e_i - P_{i-1}e_i\|_{\mathcal{V}}^2 \leq C\tau(\lambda_i^{\max})^{-1}\|g_i\|^2. \tag{3.16}$$

By definition of e_i it follows from Lemma 3.2 1., that $\|e_i\|_{\mathcal{V}} = \|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*}$ holds. Furthermore, we have

$$\|\iota_{\mathcal{U}}(Q_{i-1}g_i)\|_{\mathcal{V}_{i-1}^*} = \|\iota_{\mathcal{U}}(Q_{i-1}V_i e_i)\|_{\mathcal{V}_{i-1}^*} = \|\iota_{\mathcal{U}}(V_{i-1}P_{i-1}e_i)\|_{\mathcal{V}_{i-1}^*} = \|P_{i-1}e_i\|_{\mathcal{V}}.$$

Again the $\mathcal{V}$-orthogonality of the operator P_{i-1} implies

$$\|e_i - P_{i-1}e_i\|_{\mathcal{V}}^2 = \|e_i\|_{\mathcal{V}}^2 + \|P_{i-1}e_i\|_{\mathcal{V}}^2 - 2(e_i, P_{i-1}e_i)_{\mathcal{V}} = \|e_i\|_{\mathcal{V}}^2 - \|P_{i-1}e_i\|_{\mathcal{V}}^2.$$

Inserting the last identity in (3.16) and using (3.14) finally yields

$$C\tau(\lambda_i^{\max})^{-1}\|g_i\|^2 \geq \|e_i\|_{\mathcal{V}}^2 - \|P_{i-1}e_i\|_{\mathcal{V}}^2 \geq (1 - \kappa_\chi^2)\|e_i\|_{\mathcal{V}_i}^2,$$

which is equivalent to the assertion. □

In the context of Example 3.1, estimate (3.16) is also often called *Approximation Property* in the literature (cf., e.g., [BS08, Sec. 6.4]). In this case, assumption (3.13) is strongly related to *elliptic regularity* and holds whenever for each $g \in L^2(\Omega)$ the variational problem

$$\text{find } w \in H_0^1(\Omega) \text{ with } (\nabla w, \nabla u) = (g, u) \quad \text{for all } u \in H_0^1(\Omega)$$

has a solution $w \in H^2(\Omega)$ that satisfies

$$|w|_{H^2(\Omega)} \leq C\|g\|_{L^2(\Omega)}. \tag{3.17}$$

For a proof see for instance [BS08, Thm. 5.4.8]. Whether elliptic regularity holds depends on the domain Ω. It is well known that it is satisfied when Ω is polygonal and convex but not for polygonal domains with reentrant corners.

3.2.2. The case without regularity

In this section we derive a result similar to Lemma 3.3 but without demanding the strong assumption (3.13). Instead, we assume that the following approximation property for the $\mathcal{U}$-orthogonal projections holds:

$$\|v - Q_i v\|^2 \leq C(\lambda_i^{\max})^{-1}\|v\|_{\mathcal{V}}^2 \quad \text{for } v \in \mathcal{V}. \tag{3.18}$$

Considering the setting of Example 3.1, in comparison to the approximation property (3.13), (3.18) holds for general Lipschitz domains $\Omega \subset \mathbb{R}^d$, $d \leq 3$, triangulated by a family of quasi-uniform meshes. This was shown for instance in [BX91, Thm. 3.2].

The error estimate remains true if both norms are "shifted". Let $g \in \mathcal{U}$, then

$$\|\iota_{\mathcal{U}}(g - Q_i g)\|_{\mathcal{V}^*} = \sup_{v \in \mathcal{V}} \frac{((I - Q_i)g, v)}{\|v\|_{\mathcal{V}}} \leq \sup_{v \in \mathcal{V}} \frac{\|g\|\,\|(I - Q_i)v\|}{\|v\|_{\mathcal{V}}} \leq \frac{C}{\sqrt{\lambda_i^{\max}}}\|g\|.$$

Here, we have used the approximation property (3.18) and that Q_i is self-adjoint as operator in $\mathcal{U}$, i.e., $(Q_i u, v) = (u, Q_i v)$ holds for all $u, v \in \mathcal{U}$, which follows directly from (3.2). This proves the next lemma:

Lemma 3.4 *From the approximation property* (3.18) *follows*

$$\|\iota_{\mathcal{U}}(g - Q_i g)\|_{\mathcal{V}^*} \leq \frac{C}{\sqrt{\lambda_i^{\max}}}\|g\| \quad \textit{for all } g \in \mathcal{U}. \tag{3.19}$$

If we consider non-smooth elements $g_i \in \mathcal{V}_i$ where $Q_{i-1} g_i = 0$, we obtain from the previous lemma that

$$\|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}^*} = \|\iota_{\mathcal{U}}(g_i - Q_{i-1} g_i)\|_{\mathcal{V}^*} \leq \frac{C}{\sqrt{\lambda_{i-1}^{\max}}}\|g_i\| = \frac{C\sqrt{\tau}}{\sqrt{\lambda_i^{\max}}}\|g_i\|. \tag{3.20}$$

On the other hand, we get for the $\mathcal{U}$-norm

$$\begin{aligned}\|g_i\| &= \sup_{u\in\mathcal{U}} \frac{(u,g_i)}{\|u\|} \leq \sup_{u\in\mathcal{U}} \frac{(Q_i u, g_i)}{\|Q_i u\|} = \sup_{u_i\in\mathcal{V}_i} \frac{(u_i,g_i)}{\|u_i\|} \leq \sqrt{\lambda_i^{\max}} \sup_{u_i\in\mathcal{V}_i} \frac{(u_i,g_i)}{\|u_i\|_{\mathcal{V}}} \\ &\leq \sqrt{\lambda_i^{\max}} \sup_{v\in\mathcal{V}} \frac{(v,g_i)}{\|v\|_{\mathcal{V}}} = \sqrt{\lambda_i^{\max}} \|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}^*},\end{aligned}$$

where we have used that $\|Q_i u\| \leq \|u\|$ for all $u \in \mathcal{U}$, which follows directly from the orthogonality of the projection (3.2). This shows that on the space of oscillatory functions the $\mathcal{U}$- and $\mathcal{V}^*$-norm are equivalent with constants that are level dependent but share the same asymptotic behaviour for $\lambda_i^{\max} \to \infty$:

$$\frac{\sqrt{\lambda_i^{\max}}}{C\sqrt{\tau}} \|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}^*} \leq \|g_i\| \leq \sqrt{\lambda_i^{\max}} \|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}^*}. \tag{3.21}$$

A similar observation leads to

Lemma 3.5 *Let* (3.8), (3.18) *and* (3.6) *be satisfied. Furthermore, let* $\kappa_\chi > 0$ *be chosen such that* $C_Q\kappa_\chi < 1$, *where* C_Q *denotes the stability constant from* (3.6). *If* $g_i \in \mathcal{V}_i$ *is an element that is not smooth, i.e.,* (3.14) *holds, then*

$$\|g_i\|^2 \geq C^{-1}\tau^{-1}(1 - C_Q\kappa_\chi)^2 \lambda_i^{\max} \|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*}^2.$$

Proof With the inverse triangle inequality, (3.7) from Lemma 3.2, and (3.14) follows

$$\begin{aligned}\|\iota_{\mathcal{U}}(g_i - Q_{i-1}g_i)\|_{\mathcal{V}^*} &\geq \|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}^*} - \|\iota_{\mathcal{U}}(Q_{i-1}g_i)\|_{\mathcal{V}^*} \geq \|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*} - C_Q\|\iota_{\mathcal{U}}(Q_{i-1}g_i)\|_{\mathcal{V}_{i-1}^*} \\ &\geq (1 - C_Q\kappa_\chi)\|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*}.\end{aligned}$$

Now the assertion follows directly from Lemma 3.4 and (3.8). □

We emphasize that in comparison to the regular case we have the stronger assumption that $C_Q\kappa_\chi < 1$, which limits the choice of κ_χ. In Section 3.3.2 we will derive a result similar to the regular case with a different choice of the stationarity measure without this restriction.

3.2.3. An abstract smoothing algorithm

We will consider smoothing algorithms for the quadratic trust-region subproblem

$$\begin{aligned}&\min_{s_i\in\mathcal{V}_i} q_i(s_i) := (s_i, g_i) + \frac{1}{2}(s_i, H_i s_i) \\ &\text{subject to } \|s_i\|_i \leq \Delta_i.\end{aligned} \tag{3.22}$$

In the following, we always assume that $H_i\colon \mathcal{V}_i \to \mathcal{V}_i$ is a linear and symmetric operator which satisfies

$$(v_i, H_i u_i) \leq C_H \lambda_i^{\max} \|v_i\| \|u_i\| \quad \text{for all } u_i, v_i \in \mathcal{V}_i \tag{3.23}$$

with a level-independent constant C_H. We note that often the stronger assumption

$$(v_i, H_i u_i) \leq C_H \|v_i\|_{\mathcal{V}} \|u_i\|_{\mathcal{V}} \quad \text{for all } u_i, v_i \in \mathcal{V}_i$$

is true, which just says that the bilinear form induced by the operator H_i is bounded. Justified by Lemma 3.3 and Lemma 3.5 we make the following assumption:

Assumption 3.1 *If $v_i \in \mathcal{V}_i$ violates the smoothness property* (2.30)*, i.e.,*

$$\chi_{i-1}(0) < \kappa_\chi \chi_i(v_i) \tag{3.24}$$

holds, then

$$\|g_i\|^2 \geq c(\kappa_\chi, \tau)^2 \lambda_i^{\max} \chi_i(v_i)^2$$

is satisfied, where $g_i \in \mathcal{V}_i$ is the representation of $h_i'(v_i)$ according to Lemma 3.1. The constant $c(\kappa_\chi, \tau) > 0$ must be level-independent but could depend on κ_χ and τ.

Lemma 3.6 *Let Assumption 3.1 hold. Suppose $B_i^{-1} \colon \mathcal{V}_i \to \mathcal{V}_i$ is a linear operator that satisfies*

$$(B_i^{-1} g_i, H_i B_i^{-1} g_i) \leq \theta (g_i, B_i^{-1} g_i) \tag{3.25a}$$

with $\theta \in (0, 2)$,

$$(g_i, B_i^{-1} g_i) \geq C^{-1} (\lambda_i^{\max})^{-1} \|g_i\|^2 \tag{3.25b}$$

and

$$\frac{(g_i, B_i^{-1} g_i)}{\|B_i^{-1} g_i\|_i} \geq C^{-1} (\lambda_i^{\max})^{-1/2} \|g_i\|, \tag{3.25c}$$

where $\|\cdot\|_i$ denotes the trust-region norm on level i. Then $s_i = -t B_i^{-1} g_i$ with stepsize

$$t = \begin{cases} \min\{1, \Delta_i / \|B_i^{-1} g_i\|_i\} & \text{if } (B_i^{-1} g_i, H_i B_i^{-1} g_i) > 0, \\ \Delta_i / \|B_i^{-1} g_i\|_i & \text{otherwise} \end{cases}$$

is a feasible step of the trust-region subproblem (3.22)*. Moreover, if g_i is not smooth in the sense that* (3.24) *holds, then*

$$-q_i(s_i) \geq C^{-1} c(\kappa_\chi, \tau)(1 - \theta/2) \chi_i(v_i) \min\left\{\Delta_i,\, c(\kappa_\chi, \tau) \chi_i(v_i)\right\}$$

is satisfied for the predicted reduction of the step s_i.

Proof The feasibility of s_i follows straightforwardly from the definition of t.

Inserting s_i in $-q_i$ yields

$$-q_i(-t B_i^{-1} g_i) = -\frac{t^2}{2} (B_i^{-1} g_i, H_i B_i^{-1} g_i)_i + t (g_i, B_i^{-1} g_i).$$

If $(B_i^{-1} g_i, H_i B_i^{-1} g_i) \leq 0$, it follows from the choice of the stepsize t and (3.25c):

$$-q_i(-t B_i^{-1} g_i) \geq t (g_i, B_i^{-1} g_i) = \Delta_i \frac{(g_i, B_i^{-1} g_i)}{\|B_i^{-1} g_i\|_i} \geq C^{-1} (\lambda_i^{\max})^{-1/2} \Delta_i \|g_i\|.$$

By Assumption 3.1 and $\theta \geq 0$, we further obtain

$$-q_i(-tB_i^{-1}g_i) \geq C^{-1}\Delta_i c(\kappa_\chi, \tau)\chi_i(v_i) \geq C^{-1}\Delta_i c(\kappa_\chi, \tau)(1-\theta/2)\chi_i(v_i).$$

Hence, in this case the assertion is valid.

Let us now assume that $(B_i^{-1}g_i, H_i B_i^{-1}g_i) > 0$. Then from (3.25a) and $t \leq 1$ we infer

$$-q_i(-tB_i^{-1}g_i) \geq -t^2\frac{\theta}{2}(g_i, B_i^{-1}g_i) + t(g_i, B_i^{-1}g_i) \geq t(1-\theta/2)(g_i, B_i^{-1}g_i). \tag{3.26}$$

For a full step ($t = 1$) we obtain from (3.25b) and Assumption 3.1:

$$-q_i(-B_i^{-1}g_i) \geq C^{-1}(1-\theta/2)(\lambda_i^{\max})^{-1}\|g_i\|^2 \geq C^{-1}c(\kappa_\chi, \tau)^2(1-\theta/2)\chi_i(v_i)^2.$$

On the other hand, if $t = \Delta_i/\|B_i^{-1}g_i\|_i$, the full step is not feasible and instead we stop at the boundary of the trust region. From (3.26), (3.25c) and Assumption 3.1 follows:

$$\begin{aligned} -q_i(-tB_i^{-1}g_i) &\geq C^{-1}\Delta_i(1-\theta/2)(\lambda_i^{\max})^{-1/2}\|g_i\| \\ &\geq C^{-1}\Delta_i c(\kappa_\chi, \tau)(1-\theta/2)\chi_i(v_i). \end{aligned}$$

Taking the minimum of the estimates completes the proof. □

The choice of the *smoothing operator* B_i^{-1} is crucial. A simple example is the following operator which returns a steepest descent step. If the curvature of q_i in gradient direction is positive, the step that minimizes (3.22) neglecting the trust-region condition in direction $-g_i$ is given by

$$s_i = -\underbrace{\frac{\|g_i\|^2}{(g_i, H_i g_i)}}_{=:\omega_i} g_i. \tag{3.27}$$

If $(g_i, H_i g_i) \leq 0$, the quadratic function q_i is not bounded from below in direction $-g_i$, and as a consequence, the step that achieves the maximum descent lies on the boundary of the trust region. The next lemma shows that an operator B_i^{-1} based on this considerations satisfies the assumptions of Lemma 3.6.

Lemma 3.7 *Let*

$$\|v_i\|_i \leq C\sqrt{\lambda_i^{\max}}\|v_i\| \tag{3.28}$$

and (3.23) *hold. Then the operator*

$$B_i^{-1} = \begin{cases} \omega_i I_i & \text{if } (g_i, H_i g_i) > 0, \\ I_i & \text{else,} \end{cases}$$

where ω_i is defined as in (3.27) *and I_i denotes the identity operator on $\mathcal{V}_i$, satisfies* (3.25a) *to* (3.25c).

PROOF We first consider the case $(g_i, H_i g_i) > 0$. From (3.23) we infer

$$(g_i, B_i^{-1} g_i) \geq \frac{\|g_i\|^4}{C_H \lambda_i^{\max} \|g_i\|^2} \geq C_H^{-1} (\lambda_i^{\max})^{-1} \|g_i\|^2,$$

which shows (3.25b) with $C = C_H$. Furthermore,

$$(B_i^{-1} g_i, H_i B_i^{-1} g_i) = \frac{\|g_i\|^4}{(g_i, H_i g_i)} = (g_i, B_i^{-1} g_i)$$

holds, which implies (3.25a) with $\theta = 1$.

Finally, from $\omega_i > 0$ and (3.28) we obtain

$$\frac{(g_i, B_i^{-1} g_i)}{\|B_i^{-1} g_i\|_i} = \frac{\|g_i\|^2}{\|g_i\|_i} \geq C^{-1} (\lambda_i^{\max})^{-1/2} \|g_i\|,$$

which shows (3.25c).

If $(g_i, H_i g_i) \leq 0$, then also $(B_i^{-1} g_i, H_i B_i^{-1} g_i) \leq 0$. Since $(g_i, B_i^{-1} g_i) = \|g_i\|^2 \geq 0$, (3.25a) is obviously true for every $\theta \in (0, 2)$. We recall that we postulated $\lambda_i^{\max} \geq 1$ and hence (3.25b) and (3.25c) are also satisfied with $C = 1$. □

Remark 3.3 The step that is induced by the operator B_i^{-1} from the last lemma happens to be just the standard *Cauchy step*. It satisfies $s_i = -t^* g_i$, where t^* is the solution of the one dimensional problem

$$\min_{t>0} q_i(-t g_i) \quad \text{subject to } t\|g_i\|_i \leq \Delta_i.$$

If we choose instead $\omega_i = \beta \lambda_i^{-1}$ in (3.27) with $\beta \in (0, 2)$ and λ_i as the maximal eigenvalue of H_i, the smoother corresponds to the *Richardson method* applied to the equation $H_i s_i = -g_i$. The proof of Lemma 3.7 for this choice of ω_i is straightforward.

3.2.4. Smoothers for strictly convex trust-region subproblems

Lemma 3.7 shows that a properly scaled gradient step can achieve a level-independent Cauchy decrease. However, numerical tests suggests that this type of step is inadequate, because it does not smooth the gradient very well and hence a lot of steps are necessary before the smoothness property (2.30) is satisfied. Better results are obtained by algorithms that are based on subspace correction methods. In classical multigrid theory, these correspond to smoothers obtained by matrix splittings as for example the (block) Jacobi or Gauss-Seidel methods. We will formulate these smoothers in an abstract setting which is based on [BZ00, Xu92, Yse93].

In this section we assume that the quadratic problem is strictly convex. This is the case if and only if H_i is positive definite, i.e., $(s_i, H_i s_i) > 0$ for all $0 \neq s_i \in \mathcal{V}_i$. We show, using the theory in [BZ00], that for a large class of operators B_i^{-1} the assumptions in Lemma 3.6 are satisfied. We consider methods that minimize the function q_i either in parallel or successively over certain subspaces. This leads to two different types of smoothers: *additive* and *multiplicative* smoothers.

We assume a decomposition of the space $\mathcal{V}_i$ into l_i subspaces $\mathcal{V}_i^1, \ldots, \mathcal{V}_i^{l_i}$ such that

$$\mathcal{V}_i = \sum_{j=1}^{l_i} \mathcal{V}_i^j.$$

Note that this does not have to be a direct sum. Every element $g_i \in \mathcal{V}_i$ is represented by at least one sum of elements in $\mathcal{V}_i^j$, i.e.,

$$g_i = \sum_{j=1}^{l_i} g_i^j \quad \text{with } g_i^j \in \mathcal{V}_i^j.$$

This sum may or may not be unique. For each j we define operators $H_i^j \colon \mathcal{V}_i^j \to \mathcal{V}_i^j$ by the relation

$$(v_i^j, H_i^j u_i^j) = (v_i^j, H_i u_i^j) \quad \text{for all } u_i^j, v_i^j \in \mathcal{V}_i^j$$

and the $\mathcal{U}$-orthogonal projections $Q_i^j \colon \mathcal{V}_i \to \mathcal{V}_i^j$ by

$$(Q_i^j g_i, v_i^j) = (g_i, v_i^j) \quad \text{for all } v_i^j \in \mathcal{V}_i^j \text{ and } g_i \in \mathcal{V}_i.$$

The additive smoother is defined as the sum of the minima of q_i on each subspace $\mathcal{V}_i^j$ damped by a factor ω. It can be calculated by the following algorithm:

Algorithm 3.1 (ASmoother)
Choose a damping factor $\omega > 0$.

Step 1 Minimize $\varphi_i(s_i^j) := (Q_i^j g_i, s_i^j) + \frac{1}{2}(s_i^j, H_i^j s_i^j)$ on $\mathcal{V}_i^j$ for all $j = 1, \ldots, l_i$ and denote the solutions by s_i^{j*}.

Step 2 Set $s_i := \omega \sum_{j=1}^{l_i} s_i^{j*}$ and return with s_i.

The minimum of the function φ_i on the space $\mathcal{V}_i^j$ is attained at $s_i^{j*} = -(H_i^j)^{-1} Q_i^j g_i$. Therefore, the algorithm above corresponds to the operator defined by

$$\tilde{B}_i^{-1} := \omega \sum_{j=1}^{l_i} (H_i^j)^{-1} Q_i^j, \tag{3.29}$$

and $s_i = -\tilde{B}_i^{-1} g_i$ holds. The operators H_i^j are symmetric, which follows from the symmetry of H_i. Hence, from

$$(w_i, \tilde{B}_i^{-1} g_i) = \omega \sum_{j=1}^{l_i} (w_i, (H_i^j)^{-1} Q_i^j g_i) = \omega \sum_{j=1}^{l_i} ((H_i^j)^{-1} Q_i^j w_i, Q_i^j g_i) = (\tilde{B}_i^{-1} w_i, g_i)$$

follows the symmetry of $\tilde{B}_i^{-1}$ Furthermore, because of

$$(g_i, \tilde{B}_i^{-1} g_i) = \omega \sum_{j=1}^{l_i} (g_i, (H_i^j)^{-1} Q_i^j g_i) = \omega \sum_{i=1}^{l_i} (Q_i^j g_i, (H_i^j)^{-1} Q_i^j g_i) \geq 0,$$

it is positive semi-definite. If $(g_i, \tilde{B}_i^{-1} g_i) = 0$, it follows from the positive definiteness of H_i^j that $Q_i^j g_i = 0$ for all j. Now using $g_i = \sum_{j=1}^{l_i} g_i^j$ we obtain

$$(g_i, g_i) = \sum_{j=1}^{l_i} (g_i, g_i^j) = \sum_{j=1}^{l_i} (Q_i^j g_i, g_i^j) = 0.$$

Hence, $g_i = 0$ and the operator $\tilde{B}_i^{-1}$ is positive definite.

All subspace minimizations are independent from each other and can thus be calculated in parallel. For this reason these additive methods are often also called *parallel subspace correction methods.*

Instead of minimizing the functions independently on each subspace, it is also possible to update the step after each iteration. This leads to multiplicative smoothers:

Algorithm 3.2 (MSmoother)
Step 0 Set $s_i = 0$ and $j = 1$.

Step 1 Minimize $\varphi_i(s_i^j) := (g_i, s_i + s_i^j) + \frac{1}{2}(s_i + s_i^j, H_i(s_i + s_i^j))$ on $\mathcal{V}_i^j$ and denote the solution by s_i^{j*}.

Step 2 Update $s_i \leftarrow s_i + s_i^{j*}$. If $j < l_i$, set $j \leftarrow j + 1$ and go to Step 1, otherwise return with s_i.

Since the quadratic problems in Step 1 are strictly convex, the solutions s_i^{j*} can be expressed by

$$s_i^{j*} = -(H_i^j)^{-1} Q_i^j (g_i + H_i s_i). \tag{3.30}$$

As in the previous case, the algorithm induces a linear operator:

$$B_i^{-1} := \left[I - \prod_{j=1}^{l_i} \left(I - (H_i^{l_i - j + 1})^{-1} Q_i^{l_i - j + 1} H_i \right) \right] H_i^{-1}, \tag{3.31}$$

and $s_i = -B_i^{-1} g_i$ holds. This can be seen as follows: Define w_i^j by $w_i^j := \sum_{k=1}^{j} s_i^{k*}$ for $j = 1, \ldots, l_i$ and $w_i^0 := 0$. Then with $s_i^* := -H_i^{-1} g_i$ we obtain

$$\begin{aligned} w_i^j - s_i^* &= w_i^{j-1} + s_i^{j*} - s_i^* = w_i^{j-1} - (H_i^j)^{-1} Q_i^j (g_i + H_i w_i^{j-1}) - s_i^* \\ &= (I - (H_i^j)^{-1} Q_i^j H_i) w_i^{j-1} + (H_i^j)^{-1} Q_i^j H_i s_i^* - s_i^* = (I - (H_i^j)^{-1} Q_i^j H_i)(w_i^{j-1} - s_i^*). \end{aligned}$$

Hence, the final step s_i satisfies,

$$s_i = w_i^{l_i} - s_i^* + s_i^* = \left[I - \prod_{j=1}^{l_i} \left(I - (H_i^{l_i - j + 1})^{-1} Q_i^{l_i - j + 1} H_i \right) \right] s_i^* = -B_i^{-1} g_i.$$

The multiplicative operator B_i^{-1} is not symmetric in general. A symmetric version can be constructed by additionally minimizing q_i on the subspaces in reverse order. This leads to

$$\bar{B}_i^{-1} = \left[I - \prod_{j=1}^{l_i} \left(I - (H_i^j)^{-1} Q_i^j H_i \right) \prod_{j=1}^{l_i} \left(I - (H_i^{l_i-j+1})^{-1} Q_i^{l_i-j+1} H_i \right) \right] H_i^{-1}. \tag{3.32}$$

Another representation of the symmetric variant is

$$\bar{B}_i^{-1} = B_i^{-T} + B_i^{-1} - B_i^{-T} H_i B_i^{-1},$$

which can be shown by a straightforward calculation.

Remark 3.4 For the following theory it is not strictly necessary to solve the optimization problems on each subspace in Algorithms 3.1 and 3.2 exactly. Instead, one can replace the inverse operator $(H_i^j)^{-1}$ in (3.30) by an approximation R_i^j. Suppose, there is $\tilde{\theta} \in (0, 2)$ and $\tilde{\omega} > 0$ such that

$$\begin{aligned} (R_i^j v_i^j, H_i^j R_i^j v_i^j) &\leq \tilde{\theta}(v_i^j, R_i^j v_i^j) \quad \text{for all } v_i^j \in \mathcal{V}_i^j, \\ (v_i^j, \bar{R}_i^j v_i^j) &\geq \frac{\tilde{\omega}}{\lambda_i}(v_i^j, v_i^j) \quad \text{for all } v_i^j \in \mathcal{V}_i^j, \end{aligned}$$

where $\bar{R}_i^j = R_i^j + (R_i^j)^T - (R_i^j)^T H_i^j R_i^j$ and λ_i denotes the largest eigenvalue of H_i. Under these assumptions on R_i^j, Theorem 3.1 can also be proven (cf. [BZ00, Thm. 8.3, Thm. 8.4]). As a simple example consider $R_i^j = \theta^j (H_i^j)^{-1}$, $\theta^j \in (0, 2)$, which clearly satisfies the assumptions. This allows us to use SOR (*successive overrelaxation*) type smoothers.

Example 3.2 A simple but important example is the direct decomposition of $\mathcal{V}_i$ into the one dimensional spaces spanned by the $l_i = n_i$ basis functions ϕ_i^j, i.e., setting $\mathcal{V}_i^j = \{\alpha \phi_i^j \mid \alpha \in \mathbb{R}\}$. In this case the operators H_i^j and Q_i^j are given by

$$H_i^j v_i^j = \frac{(\phi_i^j, H_i \phi_i^j)}{(\phi_i^j, \phi_i^j)} v_i^j \qquad \text{and} \qquad Q_i^j g_i = \frac{(\phi_i^j, g_i)}{(\phi_i^j, \phi_i^j)} \phi_i^j. \tag{3.33}$$

The additive smoother becomes

$$\tilde{B}_i^{-1} g_i = \omega \sum_{j=1}^{l_i} \frac{(g_i, \phi_i^j)}{(\phi_i^j, H_i \phi_i^j)} \phi_i^j.$$

Let us assume that we have representations $\underset{\approx}{H}_i \in \mathbb{R}^{n_i \times n_i}$ of H_i and $\underset{\sim}{g}_i \in \mathbb{R}^{n_i}$ of g_i, which have the entries $\underset{\approx}{H}_i^{jk} = (\phi_i^j, H_i \phi_i^k)$ and $\underset{\sim}{g}_i^j = (g_i, \phi_i^j)$. These are the typical representations when using finite element discretizations (cf. Section 3.4 for more details). Using the additive smoother, we get for the j-th entry of the coefficient vector $\tilde{s}_i^j = -\omega \underset{\sim}{g}_i^j / \underset{\approx}{H}_i^{jj}$ and thus $\tilde{s}_i = -\omega \operatorname{Diag}(\underset{\approx}{H}_i)^{-1} \underset{\sim}{g}_i$. This is exactly one damped Jacobi iteration applied to the linear optimality system $\underset{\approx}{H}_i \tilde{s}_i = -\underset{\sim}{g}_i$. In a similar way, the multiplicative smoother is connected to a Gauss-Seidel algorithm, or to a symmetric Gauss-Seidel algorithm when using the symmetric variant. From an optimization point of view we minimize the quadratic function successively along the coordinate directions. This is also known as *sequential coordinate minimization*. More details on the classical algorithms can be found, e.g., in [Var62] or [Saa03]. It should be noted that the effort to calculate one iteration of the multiplicative algorithm for this decomposition is of order of a single matrix-vector product.

Not every space decomposition leads to smoothers that satisfy the assumptions from Lemma 3.6. For this we need to impose certain requirements. Let the matrix γ_1 be defined by

$$\gamma_1^{jk} = \begin{cases} 0 & \text{if } (v_i^j, H_i v_i^k) = 0 \text{ for all } v_i^j \in \mathcal{V}_i^j,\ v_i^k \in \mathcal{V}_i^k, \\ 1 & \text{otherwise.} \end{cases} \tag{3.34}$$

We assume that there exists a constant $\nu_1 \geq 1$, independent of i, such that

$$\|\gamma_1\|_\infty \leq \nu_1. \tag{3.35}$$

This condition says that, independent of the level, only a fixed number of subspaces are not orthogonal with respect to the inner product induced by the operator H_i. In many cases the number ν_1 is small compared to the number of subspaces. Note that if $\nu_1 = 1$, we have an orthogonal decomposition of $\mathcal{V}_i$ and one iteration of Algorithm 3.1 or 3.2 returns a step s_i that exactly minimizes q_i.

The second assumption is that for every $g_i \in \mathcal{V}_i$ there exists a decomposition $g_i = \sum_{j=1}^{l_i} g_i^j$, $g_i^j \in \mathcal{V}_i^j$, such that

$$\sum_{j=1}^{l_i} \|g_i^j\|^2 \leq C \|g_i\|^2 \tag{3.36}$$

with a constant C independent of i.

Under these two assumptions, the following theorem can be proven:

Theorem 3.1 *Let $\{\mathcal{V}_i^j\}_{j=1}^{l_i}$ be a decomposition of $\mathcal{V}_i$ such that* (3.35) *and* (3.36) *are satisfied. Then it holds:*

1. *The additive smoother $\tilde{B}_i^{-1}$, defined by* (3.29)*, satisfies* (3.25a) *and* (3.25b) *for $\omega = \theta/\nu_1$.*
2. *The smoother $\bar{B}_i^{-1}$, defined by* (3.32)*, satisfies assumptions* (3.25a) *and* (3.25b)*.*

Proof Instead of (3.25b), we show

$$(g_i, B_i^{-1} g_i) \geq C \lambda_i^{-1} \|g_i\|^2 \tag{3.37}$$

where λ_i denotes the largest eigenvalue of H_i. The estimate (3.37) implies (3.25b) because of $\lambda_i \leq C_H \lambda_i^{\max}$, which follows from (3.23):

$$\lambda_i = \sup_{v_i \in \mathcal{V}_i} \frac{(v_i, H_i v_i)}{\|v_i\|^2} \leq C_H \lambda_i^{\max}.$$

Under the stated assumptions, for the additive smoother (3.25a) follows directly from Theorem 8.1, and (3.37) from Theorem 8.7 in [BZ00].

In Theorem 8.2 in [BZ00] it is shown that (3.25a) is satisfied for the multiplicative smoother and (3.37) for the symmetric smoother provided that assumptions (3.35) and (3.36) are satisfied. We can formulate the symmetric multiplicative smoother (3.32) by the definition of the multiplicative smoother on a new decomposition of $\mathcal{V}_i$ into $2l_i$ subspaces, with $\bar{\mathcal{V}}_i^j = \mathcal{V}_i^j$ for $j \leq l_i$ and $\bar{\mathcal{V}}_i^j = \mathcal{V}_i^{2l_i+1-j}$ for $j > l_i$. This decomposition satisfies assumption (3.35) with $\bar{\nu}_1 \leq 2\nu_1$ and (3.36) with the same constant C. Hence, (3.25a) holds also in the symmetric case. □

Whether assumption (3.25c) is satisfied depends also on the choice of the trust-region norm $\|\cdot\|_i$. The nnext lemma shows that if the trust-region norm is not stronger than the norm induced by the operator H_i, (3.25c) holds without additional assumptions. This for instance is the case when $\|\cdot\|_i = \|\cdot\|_{\mathcal{V}}$ and the norm induced by H_i is equivalent to $\|\cdot\|_{\mathcal{V}}$: There exists an $\alpha > 0$ independent of i such that

$$(s_i, H_i s_i) \geq \alpha \|s_i\|_{\mathcal{V}}^2 \quad \text{for all } s_i \in \mathcal{V}_i.$$

The last assumption is satisfied for example if H_i is a suitable discretization of an elliptic operator on $\mathcal{V}$, e.g., the negative Laplace operator.

Lemma 3.8 *Let $\|g_i\|_i \leq C\sqrt{(g_i, H_i g_i)}$ for all $g_i \in \mathcal{V}_i$. If B_i^{-1} satisfies* (3.25b) *and* (3.25a)*, then condition* (3.25c) *holds as well.*

Proof

$$\frac{(g_i, B_i^{-1} g_i)}{\|B_i^{-1} g_i\|_i} \geq C^{-1} \frac{(g_i, B_i^{-1} g_i)}{\sqrt{(B_i^{-1} g_i, H_i B_i^{-1} g_i)}} \geq \frac{C^{-1}}{\sqrt{\theta}} \sqrt{(g_i, B_i^{-1} g_i)} \geq \frac{C^{-1}}{\sqrt{\theta}} (\lambda_i^{\max})^{-1/2} \|g_i\|. \qquad \square$$

If the trust-region norm only satisfies

$$\|s_i\|_i \leq C\sqrt{\lambda_i}\|s_i\| \quad \text{for all } s_i \in \mathcal{V}_i, \tag{3.38}$$

where λ_i denotes the largest eigenvalue of H_i, we can show (3.25c) under a stronger condition on the decomposition of $\mathcal{V}$. For this we define similar to γ_1 a matrix γ_0 with entries

$$\gamma_0^{jk} = \begin{cases} 0 & \text{if } (v_i^j, v_i^k) = 0 \text{ for all } v_i^j \in \mathcal{V}_i^j,\ v_i^k \in \mathcal{V}_i^k, \\ 1 & \text{otherwise.} \end{cases}$$

We demand

$$\|\gamma_0\|_\infty \leq \nu_0, \tag{3.39}$$

with a constant ν_0 independent of i.

Lemma 3.9 *Let the space decomposition of $\mathcal{V}_i$ satisfy* (3.35)*,* (3.36)*,* (3.38)*,* (3.39) *and*

$$\frac{\lambda_i}{C}\|v_i^j\|^2 \leq (v_i^j, H_i^j v_i^j) \leq C\lambda_i \|v_i^j\|^2 \quad \text{for all } v_i^j \in \mathcal{V}_i^j \text{ and } j = 1, \ldots, l_i. \tag{3.40}$$

Then both the operator $B_i^{-1} = \tilde{B}_i^{-1}$ defined by (3.29) *and $B_i^{-1} = \bar{B}_i^{-1}$ defined by* (3.32) *satisfy* (3.25c)*.*

Proof By Theorem 8.8 in [BZ00] it follows that under the stated assumptions

$$(v_i, B_i^{-1} v_i) \leq C\lambda_i^{-1}\|v_i\|^2 \quad \text{for all } v_i \in \mathcal{V}_i \tag{3.41}$$

holds for $B_i^{-1} = \tilde{B}_i^{-1}$ and $B_i^{-1} = \bar{B}_i^{-1}$. From (3.38) follows $\|B_i^{-1} g_i\|_i^2 \leq C\lambda_i \|B_i^{-1} g_i\|^2$. Since B_i^{-1} is symmetric and positive definite, we obtain from (3.41):

$$\|B_i^{-1} g_i\|_i^2 \leq C\lambda_i (B_i^{-1/2} g_i, B_i^{-1} B_i^{-1/2} g_i) \leq C(g_i, B_i^{-1} g_i).$$

Hence, from (3.25b) it follows

$$\frac{(g_i, B_i^{-1} g_i)}{\|B_i^{-1} g_i\|_i} \geq C^{-1} \frac{(g_i, B_i^{-1} g_i)}{\sqrt{(g_i, B_i^{-1} g_i)}} \geq \frac{C^{-1}}{\sqrt{\lambda_i^{\max}}} \|g_i\|.$$

□

Remark 3.5 The statement of Lemma 3.9 stays true if the operators $(H_i^j)^{-1}$ are replaced by approximations R_i^j as defined in Remark 3.4 in the definition of $\tilde{B}_i^{-1}$ and $\bar{B}_i^{-1}$ and assumption (3.40).

Example 3.3 The decomposition from Example 3.2 satisfies (3.35) when the number of non-zero entries in each row of the stiffness matrix $(\phi_i^j, H_i \phi_i^k)_{jk}$ and the mass matrix $(\phi_i^j, \phi_i^k)_{jk}$ is bounded independently of i. This is true in the majority of cases when using finite elements for the discretization, since the support of the nodal basis functions is bounded to a small number of simplices. To show (3.36), we assume that the Euclidean norm of the coefficient vectors $\tilde{v}_i$ of an element $v_i \in \mathcal{V}_i$ fulfills

$$\frac{1}{C} c_i^d \|v_i\|^2 \leq \|\tilde{v}_i\|_2^2 \leq C c_i^d \|v_i\|^2, \quad v_i = \sum_{j=1}^{n_i} \tilde{v}_i^j \phi_i^j, \tag{3.42}$$

where c_i^d is a level-dependent constant. In the setting of Example 3.1 this is a well-known fact with $c_i^d = h_i^{-d}$. It follows from the shape regularity of the triangulation, cf., e.g., [Bra07, Thm. 2.5]. With (3.42) we can estimate

$$\sum_{j=1}^{l_i} \|\tilde{v}_i^j \phi_i^j\|^2 = \sum_{j=1}^{l_i} |\tilde{v}_i^j|^2 \|\phi_i^j\|^2 \leq C (c_i^d)^{-1} \sum_{j=1}^{l_i} |\tilde{v}_i^j|^2 \leq C \|v_i\|^2. \tag{3.43}$$

The lower bound in (3.40) is satisfied for instance if the norm induced by H_i is equivalent to $\|\cdot\|_{\mathcal{V}}$ and $\|\phi_i^j\|_{\mathcal{V}}^2 \geq C^{-1} \lambda_i^{\max} \|\phi_i^j\|^2$ holds. The last inequality says that the nodal basis functions are not completely smooth but have a fixed and level independent non-smooth part. If the estimate $\|\phi_i^j\| \leq C \|\phi_i^j - Q_{i-1} \phi_i^j\|$ is satisfied, it follows directly from the approximation property (3.18).

3.2.5. A smoother for non-convex problems

The techniques used in the convex case cannot be transferred one-to-one to the non-convex case. One reason is that in the proofs a Cauchy-Schwarz type inequality for the H_i inner product is heavily used, which does not hold in the non-convex case. Moreover, the following simple example in $\mathbb{R}^2$ shows that we cannot expect a sufficient minimum decrease for general subspace minimization algorithms like Algorithm 3.2:

Example 3.4 Let $\varepsilon \in (0, 1)$ and $q \colon \mathbb{R}^2 \to \mathbb{R}$,

$$q(x) = g^T x + \frac{1}{2} x^T H x := \begin{pmatrix} \varepsilon \\ 1 \end{pmatrix}^T x + \frac{1}{2} x^T \begin{pmatrix} \varepsilon & 1 \\ 1 & 1 \end{pmatrix} x.$$

Independent of ε, the eigenvalues of H are in the interval

$$[(1-\sqrt{5})/2,\ 2]$$

and therefore (3.23) is satisfied with $C_H = 2$ for the choice $\|\cdot\| = \|\cdot\|_{\mathcal{V}} = \|\cdot\|_2$. We apply the multiplicative subspace correction algorithm (Algorithm 3.2) to q where we use the decomposition of $\mathbb{R}^2$ into the subspaces which are spanned by the unit vectors $e_1 = (1,0)^T$ and $e_2 = (0,1)^T$. This decomposition satisfies (3.35) and (3.36). We assume that the trust region is large enough to not influence the step that we calculate in the following.

- We start with $s = 0$ and $j = 1$. In the first minimization in Step 1 we obtain the solution $s^{1*} = -\frac{g^T e_1}{e_1^T H e_1} e_1 = -e_1$. The update in Step 2 yields $s = -e_1$.
- Since $\nabla q(s) = 0$, s is a stationary point. The curvature in direction e_2 is positive and therefore $t = 0$ is the global minimum of $t \mapsto q(s + te_2)$. So Algorithm 3.2 returns with the step $s = -e_1$.
- The descent of this step, however, is $q(0) - q(s) = 0 + \varepsilon - \frac{\varepsilon}{2} = \frac{\varepsilon}{2} \leq \frac{\varepsilon}{2}\|g\|_2^2$. Therefore, the descent becomes arbitrary small for $\varepsilon \to 0$ and we cannot guarantee a minimum decrease that only depends on $\|g\|_2$ and $\|H\|_2$.

We have already seen that the steepest descent step achieves level-independent Cauchy decrease even in the non-convex case. The goal of this section is to establish a smoothing algorithm that is more similar to the classical additive and multiplicative smoothers introduced in the last section. For this, we assume that we have a decomposition of $\mathcal{V}_i$ into subspaces $\mathcal{V}_i^j \subset \mathcal{V}_i$, $j = 1, \ldots, p$, where each subspace is the linear span of l_i^j basis functions

$$\Phi_i^j := \{\phi_i^{j1}, \phi_i^{j2}, \ldots, \phi_i^{jl_i^j}\} \subset \{\phi_i^j, |, j = 1, \ldots, n_i\}.$$

Furthermore, we suppose that the basis functions in Φ_i^j are pairwise H_i-orthogonal, i.e, for all $j = 1, \ldots, p$ it holds

$$(\phi_i^j, H_i \phi_i^{j'}) = 0 \text{ for } \phi_i^j, \phi_i^{j'} \in \Phi_i^j \text{ with } \phi_i^j \neq \phi_i^{j'}. \tag{3.44}$$

We stress that the number p is assumed to be level-independent.

In a typical finite element setting the support of the nodal basis functions are small which leads to a sparse stiffness matrix, i.e., for a fixed j, $(\phi_i^j, H_i \phi_i^k) \neq 0$ only for a small number of different k. This number does not depend on the meshsize of the triangulation (cf. also Example 3.3) and is level-independent for shape-regular triangulations. By graph coloring arguments it follows that in this case a decomposition that satisfies (3.44) exists (cf. Section 3.4.1 for more details). An example is given in Figure 3.3, where we assume that the support of each nodal basis function consists only of the triangles surrounding the node. This is the case for piecewise linear finite elements. The supports of the nodal basis functions belonging to the same color are disjunct and hence (3.44) with $p = 4$ is satisfied. Note that in this case the functions in Φ_i^j are also $\mathcal{U}$-orthogonal. Another typical example is the red-black or checkerboard coloring, which can be used in the finite differences setting of Example 2.2. Here the grid is divided into red and black points (like on a checkerboard) and each unit vector corresponding to a red resp. black point is

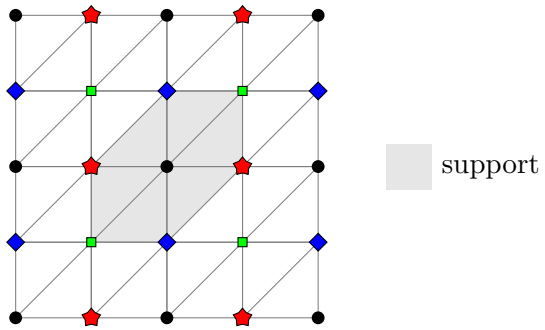

Figure 3.3.: Example coloring of a triangulation with four colors

independent in the sense of (3.44). More details about this classical coloring in the context of finite differences can be found for instance in [TOS01].

Using this decomposition we consider the following *partial successive subspace correction algorithm*:

Algorithm 3.3 (PSSC ($\{\Phi_i^j\}_j, H_i, g_i$))
Choose constants $\alpha > 0$, $\theta \in (0, 2)$ and $C_\alpha \geq \alpha$.

Step 0 For every $j \in \{1, \ldots, p\}$ calculate $\Theta^j = \sum_{\phi_i^j \in \Phi_i^j} (\phi_i^j, g_i)^2$. Define the ordered set of indices $J = (J^1, \ldots, J^p)$ such that $\Theta^{J^k} \geq \Theta^{J^{k+1}}$ for $k = 1, \ldots, p-1$. Set $k = 1$, $j = J^1$ and $s_i^0 = 0$.

Step 1 For every element $\phi_i^j \in \Phi_i^j$ calculate the step size

$$t_{\phi_i^j} = -(g_i + H_i s_i^{k-1}, \phi_i^j) / w(\phi_i^j, H_i)$$

where

$$w(\phi_i^j, H_i) := \begin{cases} (\phi_i^j, H_i \phi_i^j) & \text{if } (\phi_i^j, H_i \phi_i^j) > \alpha \lambda_i^{\max} \|\phi_i^j\|^2, \\ C_\alpha \lambda_i^{\max} \|\phi_i^j\|^2 & \text{else.} \end{cases}$$

Step 2 Set $s_i^k = s_i^{k-1} + \theta \sum_{\phi_i^j \in \Phi_i^j} t_{\phi_i^j} \phi_i^j$.

Step 3 If $\|s_i^k\|_i \leq \Delta_i$ and $k < p$, set $k \leftarrow k+1$, $j = J^k$ and go to Step 1. If $\|s_i^k\|_i \leq \Delta_i$ and $k = p$, return with s_i^k.

Step 4 Set $\hat{s}_i^k = s_i^{k-1} + t(s_i^k - s_i^{k-1})$ with the maximal stepsize $0 < t \leq 1$ such that $\|\hat{s}_i^k\|_i \leq \Delta_i$ holds. If $-q_i(\hat{s}_i^k) \geq -q_i(s_i^{k-1})$, return with $\hat{s}_i^k$, otherwise with s_i^{k-1}.

The main idea of the algorithm is to identify a partition where we could expect a good descent. We then make a step in each basis direction in this partition as in the additive smoother. If $(\phi_i^j, H_i \phi_i^j) > \alpha \|\phi_i^j\|_{\mathcal{V}}^2$, the step length $t_{\phi_i^j}$ is chosen such that $t_{\phi_i^j} \phi_i^j$ minimizes the quadratic function in the direction ϕ_i^j, i.e., solves $\min_t \varphi_i^j(t) := q_i(s_i^{k-1} + t\phi_i^j)$. In the other case, the curvature of q_i

in this direction is small, or even negative. The algorithm exploits this fact, but it is necessary to limit the length of the step in this direction. Since the elements of Φ_i^j are H_i-orthogonal, the optimizations along these are independent of each other and we can successfully handle the case where the curvature is negative. Because of the ordering of the partition, the step $\hat{s}_i^1$ already achieves (under suitable assumptions) enough descent to show the fraction of Cauchy decrease condition.

To show that a step calculated by Algorithm 3.3 satisfies the Cauchy decrease condition, we need another assumption similar to (3.39): For each $j = 1, \dots, p$ define the matrix $\gamma_{0,j} \in \{0,1\}^{l_i^j \times l_i^j}$ by

$$\gamma_{0,j}^{kk'} := \begin{cases} 0 & \text{if } (\phi_i^{jk}, \phi_i^{jk'}) = 0, \\ 1 & \text{otherwise.} \end{cases}$$

We assume that there is a level-independent constant ν_0 with

$$\nu_0 \geq \|\gamma_{0,j}\|_\infty \quad \text{for all } j = 1, \dots, p. \tag{3.45}$$

This is a rather weak assumption, which is satisfied with $\nu_0 = 1$ for instance if the basis functions in each set Φ_i^j are also $\mathcal{U}$-orthogonal. Furthermore, if (3.39) is satisfied for the decomposition of $\mathcal{V}_i$ into the spaces spanned by the nodal basis functions, then (3.45) holds as well with $\hat{\nu}_0 = \nu_0$. The following Cauchy-Schwarz type inequality, which was similarly proven in [BZ00], is the main reason for this assumption:

Lemma 3.10 *Let X be an inner product space and $v_i, u_i \in X$, $i = 1, \dots, n$. Define $\gamma \in \{0,1\}^{n \times n}$ by*

$$\gamma^{ij} = \begin{cases} 0 & \text{if } (v_i, u_j)_X = 0, \\ 1 & \text{else.} \end{cases}$$

The following estimate holds with $\nu = \max\{\|\gamma\|_\infty, \|\gamma\|_1\}$:

$$\sum_{i,j=1}^n |(v_i, u_j)_X| \leq \nu \sqrt{\sum_{i=1}^n (v_i, v_i)_X} \sqrt{\sum_{j=1}^n (u_i, u_i)_X}.$$

If γ is symmetric, we have $\nu = \|\gamma\|_\infty$.

Proof Note that $\sum_{j=1}^n \gamma^{ij} \leq \nu$ for all i and similar $\sum_{i=1}^n \gamma^{ij} \leq \nu$ for all j. We set $\|\cdot\|_X = \sqrt{(\cdot,\cdot)_X}$. Using the Cauchy-Schwarz inequality we obtain

$$\begin{aligned} \sum_{i,j=1}^n |(v_i, u_j)_X| &\leq \sum_{i,j=1}^n \gamma^{ij} \|v_i\|_X \|u_j\|_X \leq \sqrt{\sum_{i,j=1}^n \gamma^{ij} \|v_i\|_X^2} \sqrt{\sum_{i,j=1}^n \gamma^{ij} \|u_j\|_X^2} \\ &\leq \sqrt{\nu \sum_{i=1}^n \|v_i\|_X^2} \sqrt{\nu \sum_{j=1}^n \|u_j\|_X^2} = \nu \sqrt{\sum_{i=1}^n (v_i, v_i)_X} \sqrt{\sum_{j=1}^n (u_i, u_i)_X}. \end{aligned}$$

□

For the following theorem, we suppose that (3.42) holds, i.e.,

$$\frac{1}{C} c_i^d \|v_i\|^2 \leq \|\tilde{v}_i\|_2^2 \leq C c_i^d \|v_i\|^2 \quad \text{for all } v_i = \sum_{j=1}^{n_i} \tilde{v}_i^j \phi_i^j$$

is satisfied. A simple consequence is the following upper bound on the $\mathcal{V}$-norm of the basis functions:

$$\|\phi_i^j\|_{\mathcal{V}}^2 \leq \lambda_i^{\max}\|\phi_i^j\|^2 \leq \frac{C}{c_i^d}\lambda_i^{\max}. \tag{3.46}$$

Furthermore, from

$$\|v_i\|^2 = (v_i, \sum_{m=1}^{n_i} \tilde{v}_i^m \phi_i^m) = \sum_{m=1}^{n_i} \tilde{v}_i^m (v_i, \phi_i^m) \leq \|\tilde{v}_i\|_2 \sqrt{\sum_{m=1}^{n_i} (v_i, \phi_i^m)^2} \leq \sqrt{C c_i^d}\|v_i\| \sqrt{\sum_{m=1}^{n_i} (v_i, \phi_i^m)^2},$$

we infer

$$\sum_{m=1}^{n_i} (v_i, \phi_i^m)^2 \geq \frac{1}{C c_i^d}\|v_i\|^2. \tag{3.47}$$

The next theorem shows that a step generated by Algorithm 3.3 achieves a level-independent Cauchy decrease, when the gradient is not smooth.

Theorem 3.2 *Let assumptions* (3.23), (3.8), (3.18), (3.42) *and* (3.45) *hold. Furthermore, assume that* $C_\alpha \leq C_H$ *and*

$$\|v_i\|_i \leq C\sqrt{\lambda_i^{\max}}\|v_i\| \quad \textit{for all } v_i \in \mathcal{V}_i. \tag{3.48}$$

Then the step $\hat{s}_i$ *generated by Algorithm 3.3 is feasible for the trust-region subproblem. Moreover, if Assumption 3.1 and* (3.14) *hold, the predicted reduction of the step satisfies*

$$-q_i(\hat{s}_i) \geq C^{-1} p^{-1/2} \frac{2\theta - \theta^2}{2C_H} c(\kappa_\chi, \tau)\chi_i(v_i) \min\left\{ p^{-1/2} c(\kappa_\chi, \tau)\chi_i(v_i), \frac{\alpha}{\theta\sqrt{\hat{\nu}_0}}\Delta_i \right\}$$

with a level-independent constant $C > 0$.

Proof The feasibility follows directly from the conditions in Step 3 and Step 4 of the algorithm.

To show the bound on the predicted reduction of q_i, we start by estimating the descent of the step $\hat{s}_i^1$ after the first iteration. It is given by

$$\hat{s}_i^1 := t s_i^1 = t\theta \sum_{\phi_i^j \in \Phi_i^j} t_{\phi_i^j} \phi_i^j$$

where $t := \min\{1, \Delta_i / \|s_i^1\|_i\}$. Inserting the step into the quadratic function yields

$$-q_i(\hat{s}_i^1) = t\theta \sum_{m=1}^{l_i^j} \frac{(g_i, \phi_i^{jm})^2}{w(\phi_i^{jm}, H_i)} - \frac{t^2\theta^2}{2} \left(\sum_{m=1}^{l_i^j} \frac{(g_i, \phi_i^{jm})}{w(\phi_i^{jm}, H_i)} \phi_i^{jm}, \sum_{m'=1}^{l_i^j} \frac{(g_i, \phi_i^{jm'})}{w(\phi_i^{jm'}, H_i)} H_i \phi_i^{jm'} \right).$$

Since all basis functions in Φ_i^j are H_i-orthogonal, the last expression simplifies to

$$-q_i(\hat{s}_i^1) = \sum_{m=1}^{l_i^j} \left[t\theta \frac{(g_i, \phi_i^{jm})^2}{w(\phi_i^{jm}, H_i)} - \frac{t^2\theta^2}{2} \frac{(g_i, \phi_i^{jm})^2}{w(\phi_i^{jm}, H_i)^2} \left(\phi_i^{jm}, H_i \phi_i^{jm} \right) \right].$$

Now, we look at a fixed term of the sum. From the definition of the function w and (3.23) follows

$$a_m := t\theta\frac{(g_i,\phi_i^{jm})^2}{w(\phi_i^{jm},H_i)} - \frac{t^2\theta^2}{2}\frac{(g_i,\phi_i^{jm})^2}{w(\phi_i^{jm},H_i)^2}\left(\phi_i^{jm},H_i\phi_i^{jm}\right) = \left(t\theta - \frac{t^2\theta^2}{2}\right)\frac{(g_i,\phi_i^{jm})^2}{(\phi_i^{jm},H_i\phi_i^{jm})}$$
$$\geq t\frac{2\theta-\theta^2}{2}\frac{(g_i,\phi_i^{jm})^2}{C_H\lambda_i^{\max}\|\phi_i^{jm}\|^2}$$

if $(\phi_i^{jm},H_i\phi_i^{jm}) > \alpha\lambda_i^{\max}\|\phi_i^{jm}\|^2$. In the case $(\phi_i^{jm},H_i\phi_i^{jm}) \leq \alpha\lambda_i^{\max}\|\phi_i^{jm}\|^2$, we can derive the same estimate:

$$a_m \geq t\theta\frac{(g_i,\phi_i^{jm})^2}{C_\alpha\lambda_i^{\max}\|\phi_i^{jm}\|^2} - \frac{t^2\theta^2}{2}\frac{(g_i,\phi_i^{jm})^2}{(\lambda_i^{\max})^2C_\alpha^2\|\phi_i^{jm}\|^4}\alpha\lambda_i^{\max}\|\phi_i^{jm}\|^2 \geq t\frac{2\theta-\theta^2}{2}\frac{(g_i,\phi_i^{jm})^2}{C_H\lambda_i^{\max}\|\phi_i^{jm}\|^2}.$$

Hence, we have

$$-q_i(\hat{s}_i^1) \geq t\frac{2\theta-\theta^2}{2C_H\lambda_i^{\max}}\left(\sum_{m=1}^{l_i^j}\frac{(g_i,\phi_i^{jm})^2}{\|\phi_i^{jm}\|^2}\right). \tag{3.49}$$

Using the estimates (3.46), (3.47) and $\Theta^{J^1} \geq \Theta^{J^k}$ for all $k = 1,\ldots,p$ we conclude

$$\begin{aligned}\sum_{m=1}^{l_i^j}\frac{(g_i,\phi_i^{jm})^2}{\|\phi_i^{jm}\|^2} &\geq \frac{c_i^d}{C}\Theta^{J^1} \geq \frac{c_i^d}{pC}\sum_{m=1}^{n_i}(g_i,\phi_i^m)^2 \geq C^{-1}p^{-1}\|g_i\|^2 \\ &\geq C^{-1}p^{-1}c(\kappa_\chi,\tau)^2\lambda_i^{\max}\chi_i(v_i)^2,\end{aligned} \tag{3.50}$$

where we have used Assumption 3.1 to derive the last estimate. If we make the full step, i.e., $t = 1$, we hence obtain for the predicted reduction

$$-q_i(s_i^1) \geq \frac{2\theta-\theta^2}{2C_H}C^{-1}p^{-1}c(\kappa_\chi,\tau)^2\chi_i(v_i)^2. \tag{3.51}$$

On the other hand, if $t = \Delta_i/\|s_i^1\|_i$, it follows from (3.48) that $t \geq \Delta_i/(C\sqrt{\lambda_i^{\max}}\|s_i^1\|)$ holds. Lemma 3.10 applied to $\mathcal{V}_i^j$ then yields

$$\|s_i^1\|^2 = \theta^2\sum_{m,m'}^{l_i^j}(t_{\phi_i^{jm}}\phi_i^{jm}, t_{\phi_i^{jm'}}\phi_i^{jm'}) \leq \theta^2\hat{\nu}_0\sum_{m=1}^{l_i^j}(t_{\phi_i^{jm}})^2\|\phi_i^{jm}\|^2.$$

From the definition of w it follows that $w(\phi_i^{jm},H_i) \geq \alpha\lambda_i^{\max}\|\phi_i^{jm}\|^2$ for $m = 1,\ldots,l_i^j$. Therefore,

$$\|s_i^1\|^2 \leq \theta^2\hat{\nu}_0\sum_{m=1}^{l_i^j}\frac{(g_i,\phi_i^{jm})^2}{w(\phi_i^{jm},H_i)^2}\|\phi_i^{jm}\|^2 \leq \frac{\theta^2\hat{\nu}_0}{\alpha^2\lambda_i^{\max}}\sum_{m=1}^{l_i^j}\frac{(g_i,\phi_i^{jm})^2}{\lambda_i^{\max}\|\phi_i^{jm}\|^2}.$$

Inserting t in (3.49) yields

$$-\frac{2C_H}{2\theta-\theta^2}q_i(\hat{s}_i^1) \geq \frac{\Delta_i}{C\sqrt{\lambda_i^{\max}}\|s_i^1\|}\left(\sum_{m=1}^{l_i^j}\frac{(g_i,\phi_i^{jm})^2}{\lambda_i^{\max}\|\phi_i^{jm}\|^2}\right) \geq \frac{\Delta_i\alpha}{C\theta\sqrt{\hat{\nu}_0\lambda_i^{\max}}}\sqrt{\sum_{m=1}^{l_i^j}\frac{(g_i,\phi_i^{jm})^2}{\|\phi_i^{jm}\|^2}}.$$

We can estimate the sum in the last expression by (3.50) and finally obtain

$$-q_i(\hat{s}_i^1) \geq C^{-1} p^{-1/2} \frac{2\theta - \theta^2}{2C_H} c(\kappa_\chi, \tau) \chi_i(v_i) \frac{\alpha}{\theta \sqrt{\hat{\nu}_0}} \Delta_i. \tag{3.52}$$

By taking the minimum of (3.51) and (3.52) it follows that if the algorithm returns with s_i^1 or $\hat{s}_i^1$, the assertion is true.

Let us now assume that we have just finished the $(k-1)$th iteration with a step s_i^{k-1} that satisfies $\|s_i^{k-1}\|_i < \Delta_i$. Then the next step, s_i^k, has a lower function value than s_i^{k-1}. This can be seen as follows: First note that with $\delta s := \theta \sum_{\phi_i^j \in \Phi_i^j} t_{\phi_i^j} \phi_i^j$, $j = J^k$ it follows

$$-q_i(s_i^{k-1} + \delta s) = -q_i(s_i^{k-1}) - (g_i + H_i s_i^{k-1}, \delta s) - \frac{1}{2}(\delta s, H_i \delta s).$$

By the same techniques as for the first step, we can now prove that the descent produced by the step δs for the quadratic function $q_i^{k-1}(\delta s) := (g_i + H_i s_i^{k-1}, \delta s) + (\delta s, H_i \delta s)/2$ is positive and hence $-q_i(s_i^k) = -q_i(s_i^{k-1} + \delta s) \geq -q_i(s_i^{k-1})$. The test $-q_i(\hat{s}_i^k) \geq -q_i(s_i^{k-1})$ in Step 4 of the algorithm ensures that if $\|s_i^k\|_i > \Delta_i$, the final step will at least be as good as s_i^{k-1} and therefore, by induction, it obtains at least the descent of s_i^1. This completes the proof. □

3.3. Estimating the dual norm

As we have seen in the previous section, we gain level-independent descent through a smoothing step if the gradient g_i is rough in the sense of (3.14). If, on the other hand, the gradient is smooth, a successful multilevel step also achieves a descent in the objective function that is similar to a successful smoothing step (cf. Lemma 2.9). Up to this point we have always assumed that we can check whether (3.14) is satisfied in an iteration. However, in a concrete implementation this task can be very expensive depending on the normed space $\mathcal{V}$. In our typical setting, $\mathcal{V}$ is a subspace of $H^1(\Omega)$. For example let $\mathcal{V} = H_0^1(\Omega)$, $\mathcal{V}_i \subset \mathcal{V}$ be an finite dimensional subspace and $g_i \in \mathcal{V}_i$. The value of the dual norm $\|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*}$ is equal to $|v_i|_{H^1(\Omega)}$, where $v_i \in \mathcal{V}_i$ is the solution of

$$(\nabla v_i, \nabla u_i)_{L^2(\Omega)} = (g_i, u_i)_{L^2(\Omega)} \quad \text{for all } u \in H_0^1(\Omega). \tag{3.53}$$

This follows from

$$\|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*} = \sup_{0 \neq u_i \in \mathcal{V}_i} \frac{(g_i, u_i)_{L^2(\Omega)}}{|u_i|_{H^1(\Omega)}} = \sup_{0 \neq u_i \in \mathcal{V}_i} \frac{(\nabla v_i, \nabla u_i)_{L^2(\Omega)}}{|u_i|_{H^1(\Omega)}} = |v_i|_{H^1(\Omega)},$$

where the last equality is a consequence of the Cauchy-Schwarz inequality. The solution of the discrete variational equality (3.53) is in general too expensive to calculate since the condition of the resulting linear system grows quadratically with the dimension of $\mathcal{V}_i$.

So instead of doing an exact calculation of the quotient

$$\|\iota_{\mathcal{U}}(Q_{i-1} g_i)\|_{\mathcal{V}_{i-1}^*} / \|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*}, \tag{3.54}$$

we will approximate it in a suitable way. Based on this approximation we will then present a new multilevel stationarity measure that can be used as a substitution for the dual norm of the derivative.

3.3.1. Additive multilevel preconditioner

In this section we restrict ourselves to the setting of Example 3.1. We emphasize that in this setting, besides Ω being polygonal, no additional assumption about the domain Ω was made. In particular, we do not make any regularity assumptions.

One way of estimating the dual norm of the gradient is to use additive multilevel preconditioners. The idea is to apply the operator that represents one cycle of the additive smoothing algorithm (Algorithm 3.1) using a special multilevel space decomposition. As in Example 3.2 we decompose $\mathcal{V}_i$ into one dimensional subspaces spanned by the nodal basis functions. But instead of using only ϕ_i^j, $j = 1, \dots, n_i$, we also add all basis functions of the coarser spaces $\mathcal{V}_k$, $k < i$. More precisely, we assume the decomposition

$$\mathcal{V}_i = \sum_{k=1}^{i} \sum_{j=1}^{n_k} \mathcal{V}_k^j, \quad \mathcal{V}_k^j = \{\alpha \phi_k^j \,|\, \alpha \in \mathbb{R}\}.$$

If H_i is positive definite, Algorithm 3.1 applied to this decomposition leads to the symmetric and positive definite operator $\tilde{B}_i^{-1}$ (3.29) which can be used as preconditioner for instance in a conjugate gradient (CG) algorithm. The important feature of this simple preconditioner is that the condition number of the operator $\tilde{B}_i^{-1} H_i$ is bounded level-independently in many scenarios. Moreover, even when replacing $(H_i^j)^{-1}$ in (3.29) by a suitable scaling, a level-independent condition number can be shown.

Since we want to replace the evaluation of the dual norm, we consider these method for the simple Laplace equation (3.53).

We first look at the MDS (Multilevel diagonal scaling)-method proposed in [Zha92], which is just Algorithm 3.1 applied to the multilevel nodal decomposition.

Theorem 3.3 *In the setting of Example 3.1, the* MDS preconditioner $\tilde{M}_i^{-1} \colon \mathcal{V}_i \to \mathcal{V}_i$ *defined by*

$$\tilde{M}_i^{-1} g_i := \sum_{k=1}^{i} \sum_{j=1}^{n_k} \frac{(g_i, \phi_k^j)}{(\phi_k^j, \phi_k^j)_{\mathcal{V}}} \phi_k^j$$

satisfies

$$\frac{1}{C}(v_i, V_i v_i) \leq (\tilde{M}_i^{-1} V_i v_i, V_i v_i) \leq C(v_i, V_i v_i) \quad \text{for all } v_i \in \mathcal{V}_i$$

with a constant C that is independent of i and h_i.

Proof A proof is given in [Zha92, Thm. 3.1] or [Osw94, Thm. 19]. □

From the previous theorem it follows that

$$\frac{1}{C}(g_i, V_i^{-1} g_i) \leq (g_i, \tilde{M}_i^{-1} g_i) \leq C(g_i, V_i^{-1} g_i)$$

and hence with Lemma 3.2 we get

$$\frac{1}{C} \|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*}^2 \leq (g_i, \tilde{M}_i^{-1} g_i) \leq C \|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*}^2.$$

Therefore, the norm induced by $\tilde{M}_i^{-1}$ is equivalent to the dual norm and we can replace the smoothness condition by

$$\frac{\sqrt{(Q_{i-1}g_i, \tilde{M}_{i-1}^{-1}Q_{i-1}g_i)}}{\sqrt{(g_i, \tilde{M}_i^{-1}g_i)}} = \frac{\sum_{k=1}^{i-1}\sum_{j=1}^{n_k}\frac{(Q_{i-1}g_i,\phi_k^j)^2}{(\phi_k^j,\phi_k^j)_{\mathcal{V}}}}{\sum_{k=1}^{i}\sum_{j=1}^{n_k}\frac{(g_i,\phi_k^j)^2}{(\phi_k^j,\phi_k^j)_{\mathcal{V}}}} \geq \tilde{\kappa}_\chi$$

with a suitably chosen $\tilde{\kappa}_\chi$.

A similar level-independent condition holds also for the even more simple BPX preconditioner (named after its inventors Bramble, Pasciak and Xu) presented in [BPX90]:

Theorem 3.4 *Under the assumptions of Example 3.1, the* multilevel nodal basis preconditioner $\hat{M}_i^{-1}$ *defined by*

$$\hat{M}_i^{-1}g_i := \sum_{k=1}^{i} h_k^{2-d}\sum_{j=1}^{n_k}(g_i,\phi_k^j)\phi_k^j$$

satisfies

$$\frac{1}{C}(v_i, V_i v_i) \leq (\hat{M}_i^{-1}V_i v_i, V_i v_i) \leq C(v_i, V_i v_i) \quad \text{for all } v_i \in \mathcal{V}_i,$$

with a constant C that is independent of i and h_i.

Proof There are various proofs of this theorem, cf., e.g., [Zha92, Thm. 3.1] , [Osw94, Thm. 19].□

Remark 3.6 Both preconditioners can be modified without changing the level-independent condition number, by solving the coarse grid problem exactly. This leads to

$$\tilde{M}_i^{-1}g_i = V_1^{-1}Q_1 g_i + \sum_{k=2}^{i}\sum_{j=1}^{n_k}\frac{(g_i,\phi_k^j)}{(\phi_k^j,\phi_k^j)_{\mathcal{V}}}\phi_k^j \quad \text{resp.} \quad \hat{M}_i^{-1}g_i = V_1^{-1}Q_1 g_i + \sum_{k=2}^{i} h_k^{2-d}\sum_{j=1}^{n_k}(g_i,\phi_k^j)\phi_k^j.$$

3.3.2. A multilevel stationarity measure

A natural question is whether we can directly use the multilevel norms applied to g_i, examined in the last section, as stationarity measures. Since they are equivalent to the dual norm of the derivative, they clearly induce a continuous stationarity measure in the sense of Definition 2.2. Furthermore, Assumption 3.1 is satisfied as we will show in the next lemma.

Lemma 3.11 *Assume that χ_i is a stationarity measure defined by*

$$\chi_i(v_i) := \left(\sum_{k=1}^{i}\sum_{j=1}^{n_k}\langle h_i'(v_i),\phi_k^j\rangle^2 n(\phi_k^j)\right)^{1/2}, \tag{3.55}$$

where $n\colon \mathcal{V}_i \to \mathbb{R}$ satisfies $n(\phi_i^j) \leq C(\lambda_i^{\max})^{-1}\|\phi_i^j\|^{-2}$ for all $j = 1,\dots,n_i$. Furthermore, let (3.42) *and* (3.39) *hold. Then Assumption 3.1 is satisfied with*

$$c(\kappa_\chi,\tau) = C^{-1}\nu_0^{-1}\sqrt{1-\kappa_\chi^2}.$$

PROOF By definition of the lower-level models it follows that for all $k = 1, \dots, i-1$ and $j = 1, \dots, n_k$

$$\langle h'_{i-1}(0), \phi_k^j \rangle = \langle (P_{i-1}^i)^* h'_i(v_i), \phi_k^j \rangle = \langle h'_i(v_i), \phi_k^j \rangle$$

holds which implies

$$\chi_{i-1}(0)^2 = \sum_{k=1}^{i-1} \sum_{j=1}^{n_k} \langle h'_i(v_i), \phi_k^j \rangle^2 n(\phi_k^j).$$

Suppose that $v_i \in \mathcal{V}_i$ satisfies (3.24). This means

$$\sum_{k=1}^{i} \sum_{j=1}^{n_k} \langle h'_i(v_i), \phi_k^j \rangle^2 n(\phi_k^j) - \sum_{j=1}^{n_i} \langle h'_i(v_i), \phi_i^j \rangle^2 n(\phi_i^j) < \kappa_\chi^2 \sum_{k=1}^{i} \sum_{j=1}^{n_k} \langle h'_i(v_i), \phi_k^j \rangle^2 n(\phi_k^j),$$

which is equivalent to

$$(1 - \kappa_\chi^2) \chi_i^2(v_i) < \sum_{j=1}^{n_i} \langle h'_i(v_i), \phi_i^j \rangle^2 n(\phi_i^j). \tag{3.56}$$

As usual we denote the representation of $h'_i(v_i)$ with respect to $(\cdot, \cdot)$ by g_i. Since $g_i \in \mathcal{V}_i$, there exists a coefficient vector $\tilde{g}_i$ with $g_i = \sum_{k=1}^{n_i} \tilde{g}_i^k \phi_i^k$. Using this representation and the entries of the matrix γ_0 from (3.39), we obtain

$$\begin{aligned} \sum_{j=1}^{n_i} (g_i, \phi_i^j)^2 n(\phi_i^j) &\le C(\lambda_i^{\max})^{-1} \sum_{j=1}^{n_i} \Big(\sum_{k=1}^{n_i} \gamma_0^{jk} (\tilde{g}_i^k \phi_i^k, \phi_i^j) \|\phi_i^j\|^{-1} \Big)^2 \\ &\le C(\lambda_i^{\max})^{-1} \sum_{j=1}^{n_i} \Big(\sum_{k=1}^{n_i} \gamma_0^{jk} \|\tilde{g}_i^k \phi_i^k\| \Big)^2 \end{aligned}$$

where we have used the Cauchy-Schwarz inequality in the last step. We recall that for real numbers $a_m \ge 0$

$$\Big(\sum_{m=1}^{n} a_m \Big)^2 \le n \sum_{m=1}^{n} a_m^2$$

holds, which follows directly from Jensen's inequality. Since $\|\gamma_0\|_\infty \le \nu_0$, the inner sum has at most ν_0 non-zero terms for each j. Hence, we can further deduce

$$\sum_{j=1}^{n_i} (g_i, \phi_i^j)^2 n(\phi_i^j) \le C(\lambda_i^{\max})^{-1} \nu_0 \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} \gamma_0^{jk} \|\tilde{g}_i^k \phi_i^k\|^2 \le C(\lambda_i^{\max})^{-1} \nu_0^2 \sum_{j=1}^{n_i} \|\tilde{g}_i^j \phi_i^j\|^2.$$

Using (3.43), which follows from (3.42), to estimate the last sum we finally have

$$\sum_{j=1}^{n_i} (g_i, \phi_i^j)^2 n(\phi_i^j) \le C(\lambda_i^{\max})^{-1} \nu_0^2 \|g_i\|^2.$$

Inserting this result in (3.56) yields the assertion. □

If the function n in (3.55) is given by

$$n(\phi_k^j) = (\phi_k^j, \phi_k^j)_{\mathcal{V}}^{-1} \quad \text{or} \quad n(\phi_k^j) = h_k^{2-d},$$

we obtain the preconditioners of Theorem 3.3 and 3.4. Hence, it is convenient to use these directly instead of the dual norms. Another benefit is that Assumption 3.1 is satisfied for all $\kappa_\chi \in (0,1)$ in comparison to the non-regular case, where we had to assume $\kappa_\chi < C_Q^{-1}$ (cf. Lemma 3.5).

Before we finish this section, we present a result that can be used in a practical implementation. Assume that χ_i, $i = 1, \dots, r$, are chosen as in Lemma 3.11 and that

$$\frac{1-\kappa_\chi^2}{\kappa_\chi^2} \sum_{k=s}^{i-1} \sum_{j=1}^{n_k} \langle h_i'(v_i), \phi_k^j \rangle^2 n(\phi_k^j)^2 \geq \sum_{j=1}^{n_i} \langle h_i'(v_i), \phi_i^j \rangle^2 n(\phi_i^j)^2$$

holds for some $s \in \{1, \dots, i-1\}$. Then it follows directly that

$$\chi_{i-1}(0)^2 \geq \kappa_\chi^2 \chi_i(v_i)^2$$

and hence the violation of the smoothness property. This can be used as a "quick test" for s near $i-1$.

3.4. Implementation

In this section we will give a short summary on how the smoothing algorithms and the dual norm estimates can be implemented efficiently. This is important since one smoothing iteration should be inexpensive in terms of computational costs. The same should hold for the (approximate) evaluation of the smoothness quotient (3.54). We will show that by a suitable implementation the number of floating point operations (flops) for the typical decomposition of $\mathcal{V}_i$ into the one dimensional spaces generated by the n_i basis functions (cf. Example 3.2) is of order $\mathrm{O}(n_i)$ on level i.

We assume that an element $v_i \in \mathcal{V}_i$ is represented in terms of its coefficient vector $\tilde{v}_i \in \mathbb{R}^{n_i}$ with respect to the basis $\{\phi_i^j\}_{j=1,\dots,n_i} \subset \mathcal{V}_i$, $v_i = \sum_{j=1}^{n_i} \tilde{v}_i^j \phi_i^j$. For elements that have their origin in the dual space, the canonical representation is given by their action on the basis. Hence, the gradient g_i is not represented by its coefficients, but by the vector $\underset{\sim}{g}_i \in \mathbb{R}^{n_i}$ whose entries are $\underset{\sim}{g}_i^j = (\phi_i^j, g_i)$, $j = 1, \dots, n_i$.

The operators H_i are identified by matrices $\underset{\approx}{H}_i \in \mathbb{R}^{n_i \times n_i}$ such that $(v_i, H_i u_i) = \tilde{v}_i^T \underset{\approx}{H}_i \tilde{u}_i$ is satisfied. Obviously, the matrices with elements $\underset{\approx}{H}_i^{jk} = (\phi_i^j, H_i \phi_i^k)$ have this property. Similarly, we have matrices $\underset{\approx}{V}_i$ for the operators V_i and furthermore we define the *Gram matrix* or *mass matrix* $\underset{\approx}{G}_i$ by $\underset{\approx}{G}_i^{jk} = (\phi_i^j, \phi_i^k)$. Since $(u_i, v_i) = \tilde{u}_i^T \underset{\approx}{G}_i \tilde{v}_i$, these are used to calculate the inner product and the norm on $\mathcal{U}$.

With this notation the quadratic function (3.22) can be evaluated by

$$\tilde{q}_i(\tilde{s}_i) := \tilde{s}_i^T \underset{\sim}{g}_i + \frac{1}{2} \tilde{s}_i^T \underset{\approx}{H}_i \tilde{s}_i.$$

The standard Euclidean gradient of $\tilde{q}_i$ in $\mathbb{R}^{n_i}$, $\nabla\tilde{q}_i(\tilde{s}_i) = (\frac{\partial\tilde{q}_i(\tilde{s}_i)}{\tilde{s}_i^1}, \dots, \frac{\partial\tilde{q}_i(\tilde{s}_i)}{\partial\tilde{s}_i^{n_i}})^T$, corresponds directly to the $\mathcal{U}$-representation of the Fréchet derivative of q_i, i.e., for $v_i \in \mathcal{V}_i$ with coefficient vector $\tilde{v}_i \in \mathbb{R}^{n_i}$ we have the identity

$$\langle q_i'(s_i), v_i\rangle = (\nabla_{\mathcal{U}} q_i(s_i), v_i) = \nabla\tilde{q}_i(\tilde{s}_i)^T\tilde{v}_i.$$

3.4.1. Smoothers

We will first analyze how the smoothers in this chapter and the estimates of the dual-norm can be implemented and estimate their computational complexity.

Steepest descent step

The simple step $s_i = -t\|g_i\|^2/(g_i, H_i g_i)g_i$, which corresponds to the minimization in direction of the steepest descent with t as in Lemma 3.6, is surprisingly expensive to implement. The coefficient vector $\tilde{s}_i$ of s_i is given by

$$\tilde{s}_i = -t\frac{\underset{\sim}{g}_i^T \underset{\approx}{G}_i^{-1}\underset{\sim}{g}_i}{(\underset{\approx}{G}_i^{-1}\underset{\sim}{g}_i)^T \underset{\approx}{H}_i \underset{\approx}{G}_i^{-1}\underset{\sim}{g}_i}\underset{\approx}{G}_i^{-1}\underset{\sim}{g}_i.$$

Due to this, we have to solve the linear system $\tilde{g}_i = \underset{\approx}{G}_i^{-1}\underset{\sim}{g}_i$ in each step. Although the dimension of the matrix $\underset{\approx}{G}_i$ depends on n_i, the condition number is often level-independent, for example in the setting of Example 3.1. Therefore, a simple iterative algorithm like a conjugate gradient method should give an adequate approximation after a fixed number of steps independent of n_i. If we further assume that $\underset{\approx}{H}_i$ and $\underset{\approx}{G}_i$ are sparse, i.e., the number of entries per row is bounded independently of i, we get that a good approximation of the step can be calculated in $O(n_i)$ operations. However, although the condition number of $\underset{\approx}{G}_i$ is level-independent, the approximate solution of $\underset{\approx}{G}_i\tilde{g}_i = \underset{\sim}{g}_i$ is still quite expensive.

An alternative steepest descent step

Instead of minimizing q in the direction g_i, an alternative is to search for a minimizer in direction $\sum_{j=1}^{n_i}\underset{\sim}{g}_i\phi_i^j$. This leads to the step

$$s_i = -t\frac{\|\underset{\sim}{g}_i\|_2^2}{\underset{\sim}{g}_i^T\underset{\approx}{H}_i\underset{\sim}{g}_i}\sum_{j=1}^{n_i}(g_i, \phi_i^j)\phi_i^j. \tag{3.57}$$

As usual we denote by $\|\cdot\|_2$ the Euclidean norm. The corresponding coefficient vector of the step is hence given by

$$\tilde{s}_i = -t\frac{\|\underset{\sim}{g}_i\|_2^2}{\underset{\sim}{g}_i^T\underset{\approx}{H}_i\underset{\sim}{g}_i}\underset{\sim}{g}_i$$

and can thus be calculated without inverting the Gram matrix. Under the assumption that $\underset{\approx}{H}_i$ is sparse, the evaluation needs $O(n_i)$ flops. The next lemma gives us a result similar to Lemma 3.7 for this choice.

Lemma 3.12 *Let Assumption 3.1,* (3.42) *and* (3.28) *be satisfied. Furthermore, let* B_i^{-1} *be defined by*

$$B_i^{-1} g_i := \omega_i \sum_{j=1}^{n_i} (g_i, \phi_i^j) \phi_i^j$$

with

$$\omega_i := \begin{cases} \frac{\|g_i\|_2^2}{\underset{\sim}{g}_i^T \underset{\approx}{H}_i \underset{\sim}{g}_i} & \text{if } \underset{\sim}{g}_i^T \underset{\approx}{H}_i \underset{\sim}{g}_i > 0, \\ c_i^d & \text{otherwise.} \end{cases}$$

Then the step $s_i = -t B_i^{-1} g_i$ *with*

$$t = \begin{cases} \min\{1,\, \Delta_i / \|B_i^{-1} g_i\|_i\} & \text{if } \underset{\sim}{g}_i^T \underset{\approx}{H}_i \underset{\sim}{g}_i > 0, \\ \Delta_i / \|B_i^{-1} g_i\|_i & \text{otherwise,} \end{cases}$$

fulfills

$$-q_i(s_i) \geq C^{-1} C(\kappa_\chi, \tau)(1 - \theta/2) \chi_i(v_i) \min\{\Delta_i,\, C(\kappa_\chi, \tau) \chi_i(v_i)\}.$$

Proof We will show that the operator satisfies (3.25a)–(3.25c); then the assertion follows directly from Lemma 3.6.

First, assume that $\underset{\sim}{g}_i^T \underset{\approx}{H}_i \underset{\sim}{g}_i > 0$. Since

$$(B_i^{-1} g_i, H_i B_i^{-1} g_i) = \frac{\|\underset{\sim}{g}_i\|_2^4}{(\underset{\sim}{g}_i^T \underset{\approx}{H}_i \underset{\sim}{g}_i)^2} \left(\sum_{j=1}^{n_j} \underset{\sim}{g}_i^j \phi_i^j, H_i \sum_{j=1}^{n_j} \underset{\sim}{g}_i^j \phi_i^j \right) = \frac{\|\underset{\sim}{g}_i\|_2^4}{\underset{\sim}{g}_i^T \underset{\approx}{H}_i \underset{\sim}{g}_i} = (g_i, B_i^{-1} g_i),$$

(3.25a) holds with $\theta = 1$.

Furthermore, we have

$$(g_i, B_i^{-1} g_i) = \frac{\|\underset{\sim}{g}_i\|_2^4}{\underset{\sim}{g}_i^T \underset{\approx}{H}_i \underset{\sim}{g}_i} \geq \frac{\|\underset{\sim}{g}_i\|_2^2}{\lambda^{\max}(\underset{\approx}{H}_i)},$$

where $\lambda^{\max}(\underset{\approx}{H}_i)$ denotes the largest eigenvalue of the stiffness matrix. From the definition of $\lambda_i^{\max}$ and (3.23) follows

$$\tilde{v}_i^T \underset{\approx}{H}_i \tilde{v}_i \leq C_H \lambda_i^{\max} \|v_i\|^2 \quad \text{for all } v_i = \sum_{j=1}^{n_i} \tilde{v}_i^j \phi_i^j.$$

Let $\tilde{u}_i$ be an eigenvector of $\underset{\approx}{H}_i$ to the eigenvalue $\lambda^{\max}(\underset{\approx}{H}_i)$. Then using (3.42) we obtain

$$\tilde{u}_i^T \underset{\approx}{H}_i \tilde{u}_i = \lambda^{\max}(\underset{\approx}{H}_i) \|\tilde{u}_i\|_2^2 \geq C^{-1} \lambda^{\max}(\underset{\approx}{H}_i) c_i^d \|u_i\|^2$$

and thus the following upper bound on $\lambda^{\max}(\underset{\approx}{H}_i)$:

$$\lambda^{\max}(\underset{\approx}{H}_i) \leq C C_H \frac{\lambda_i^{\max}}{c_i^d}.$$

Together with (3.47), which is a consequence of (3.42), this shows (3.25b):

$$(g_i, B_i^{-1} g_i) \geq \frac{c_i^d \|\underset{\sim}{g}_i\|_2^2}{C C_H \lambda_i^{\max}} \geq C^{-1} (\lambda_i^{\max})^{-1} \|g_i\|^2.$$

From the definition of $\underset{\approx}{G}_i$ and (3.42) we infer

$$\tilde{v}_i^T \underset{\approx}{G}_i \tilde{v}_i = \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} \tilde{v}_i^j (\phi_i^j, \phi_i^k) \tilde{v}_i^k = \|v_i\|^2 \leq C (c_i^d)^{-1} \|\tilde{v}_i\|_2^2 \quad \text{for all } \tilde{v}_i \in \mathbb{R}^{n_i}. \tag{3.58}$$

From (3.28), i.e.,

$$\|s_i\|_i \leq C \sqrt{\lambda_i^{\max}} \|s_i\| \quad \text{for all } s_i \in \mathcal{V}_i,$$

(3.58) and (3.47) it further follows that

$$\frac{(g_i, B_i^{-1} g_i)}{\|B_i^{-1} g_i\|_i} \geq \frac{1}{C \sqrt{\lambda_i^{\max}}} \frac{\|\underset{\sim}{g}_i\|_2^2}{\sqrt{\underset{\sim}{g}_i^T \underset{\approx}{G}_i \underset{\sim}{g}_i}} \geq \frac{1}{C \sqrt{\lambda_i^{\max}}} \sqrt{c_i^d} \|\underset{\sim}{g}_i\|_2 \geq \frac{1}{C \sqrt{\lambda_i^{\max}}} \|g_i\|,$$

which proves (3.25c).

Now, let $\underset{\sim}{g}_i^T \underset{\approx}{H}_i \underset{\sim}{g}_i \leq 0$. This is equivalent to $(B_i^{-1} g_i, H_i B_i^{-1} g_i) \leq 0$ and since $c_i^d > 0$, (3.25a) follows. To prove (3.25b), we use (3.47) again:

$$(g_i, B_i^{-1} g_i) = c_i^d \|\underset{\sim}{g}_i\|_2^2 \geq C^{-1} \|g_i\|^2 \geq C^{-1} (\lambda_i^{\max})^{-1} \|g_i\|^2.$$

Note that we have used $\lambda_i^{\max} \geq 1$, which holds by definition. Property (3.25c) follows as in the other case. This finishes the proof. □

Remark 3.7 The direction used in the previous lemma corresponds to an Euclidean steepest descent direction of $\tilde{q}_i$ at $\tilde{s}_i = 0$. In contrast, the direction of Lemma 3.7 is a steepest descent direction of $\tilde{q}_i$ at $\tilde{s}_i = 0$ corresponding to the inner product $(\tilde{u}_i, \tilde{v}_i)_{\underset{\approx}{G}_i} := \tilde{u}_i^T \underset{\approx}{G}_i \tilde{v}_i$.

Additive and multiplicative smoothers

The implementation of the smoothers presented in Section 3.2.4 depends on the decomposition $\{\mathcal{V}_i^j\}_j$. Let us consider the simple decomposition into the basis functions as presented in Example 3.2. In this case the additive smoother $\tilde{B}_i^{-1} g_i$ can simply be evaluated by

$$\tilde{v}_i = \omega \operatorname{Diag}(\underset{\approx}{H}_i)^{-1} \underset{\sim}{g}_i,$$

which obviously needs only $\mathrm{O}(n_i)$ operations. By $\operatorname{Diag}(\underset{\approx}{H}_i)$ we denote the matrix in $\mathbb{R}^{n_i \times n_i}$ that consists only of the diagonal entries of $\underset{\approx}{H}_i$.

For the implementation of the multiplicative smoother (Algorithm 3.2), we use the following algorithm:

Algorithm 3.4
Step 0 Set $\underset{\sim}{r}_i = \underset{\sim}{g}_i$, $\tilde{s}_i = 0$ and $j = 1$.

Step 1 Calculate $\tilde{s}_i^j = -\frac{\underset{\sim}{r}_i^j}{\underset{\approx}{H}_i^{jj}}$.

Step 2 Update $\underset{\sim}{r}_i \leftarrow \underset{\sim}{r}_i + \tilde{s}_i^j \underset{\approx}{H}_i e_i^j$, where $e_i^j \in \mathbb{R}^{n_i}$ denotes the j-th unit vector.

Step 3 If $j < n_i$, set $j \leftarrow j + 1$ and go to Step 1. Otherwise return with $\tilde{s}_i$.

To show that the algorithm indeed calculates the correct step, we use the representation (3.30) of the intermediate steps s_i^{j*}. Since we use the splitting into the spaces spanned by the basis functions, the operators H_i^j and Q_i^j are given by (3.33). Thus we obtain with $w_i^0 = 0$ and $w_i^j = \sum_{k=1}^j s_i^{k*}$:

$$s_i^{j*} = -\frac{(g_i + H_i w_i^{j-1}, \phi_i^j)}{(\phi_i^j, H_i \phi_i^j)} \phi_i^j.$$

Hence, in the corresponding coefficient vector only the j-th entry is not equal zero. This implies that the entries of the coefficient vector of the final step are simply given by

$$\tilde{s}_i^j = -\frac{(g_i + H_i w_i^{j-1}, \phi_i^j)}{(\phi_i^j, H_i \phi_i^j)} \quad \text{for } j = 1, \ldots, n_i.$$

It is left to show that in Step 1 of the algorithm $\underset{\sim}{r}_i^k = (g_i + H_i w_i^{j-1}, \phi_i^k)$ holds for all $k = 1, \ldots, n_i$. We prove this by induction. Since $w_i^0 = 0$ and $\underset{\sim}{r}_i$ is initialized by $\underset{\sim}{g}_i$ this is true for $j = 1$. Now suppose that the assumption is true for fixed j. The vector $\underset{\sim}{r}_i$ is updated in Step 2. From the induction hypothesis we infer for the corresponding element $r_i \in \mathcal{V}_i$

$$r_i = g_i + H_i w_i^{j-1} + \tilde{s}_i^j H_i \phi_i^j = g_i + H_i(w_i^{j-1} + \tilde{s}_i^j \phi_i^j) = g_i + H_i(w_i^{j-1} + s_i^{j*}). = g_i + H_i w_i^j$$

This shows that in the next iteration the assumption holds, which finishes the induction.

Assumptions (3.35) implies that the numbers of entries in each column of the matrix $\underset{\approx}{H}_i$ is bounded independently of i. Therefore, the matrix-vector product in Step 2 can be implemented with $\mathrm{O}(1)$ operations and the whole algorithm requires $\mathrm{O}(n_i)$ operations. The same is true for the symmetric variant.

Remark 3.8 In Step 2 of the algorithm it is not necessary to update the whole vector $\underset{\sim}{r}_i$. It is enough to consider only the components $\underset{\sim}{r}_i^k$ with $k = j + 1, \ldots, n_i$, which are needed for the further iterations.

Remark 3.9 Suppose the typical matrix splitting $\underset{\approx}{H}_i = \underset{\approx}{D}_i - \underset{\approx}{L}_i - \underset{\approx}{L}_i^T$ in a diagonal and a lower left triangular matrix. By a simple induction one can show that the application of Algorithm 3.4 can also be expressed by

$$\tilde{s}_i = (\underset{\approx}{D}_i - \underset{\approx}{L}_i)^{-1} \underset{\sim}{g}_i.$$

This corresponds to one iteration of the classical Gauss-Seidel algorithm applied to the linear system $\underset{\approx}{H}_i \tilde{s}_i = \underset{\sim}{g}_i$. Similarly, one iteration of the symmetric variant is given by

$$\tilde{s}_i = (\underset{\approx}{D}_i - \underset{\approx}{L}_i^T)^{-1} \underset{\approx}{D}_i (\underset{\approx}{D}_i - \underset{\approx}{L}_i)^{-1} \underset{\sim}{g}_i.$$

Graph coloring

To use Algorithm 3.3, we need to group the nodal basis functions, such that all entries in each partition are pairwise orthogonal with respect to the bilinear form $(\cdot, H_i \cdot)$. Given the matrix representation $\underset{\approx}{H}_i$ of the operator H_i, we seek index sets $I_1, \dots, I_p$ such that

$$(e_j^k)^T \underset{\approx}{H}_i e_j^{k'} = 0 \text{ for all } k, k' \in I_j,\ k \neq k'.$$

Here, $e_j^k \in \mathbb{R}^{n_i}$ denotes the unit vector, which is one at the k-th entry of I_j and zero otherwise. This property depends only on the sparsity pattern of the matrix, which in most applications does not change during the iterations since it is determined by the discretization and does not depend on the point $v_{i,k}$. In this case, we have to define these sets just once for each level when we enter it for the first time.

In order to obtain such index sets, graph coloring algorithms can be used. Then each set consists of nodes that have the same color. For this we interpret $\underset{\approx}{H}_i$ as adjacency matrix where we assume a connection between two nodes k and k' if the entry $\underset{\approx}{H}_i^{kk'}$ is not equal to zero. This is just the matrix γ_1 defined by (3.34). Consider for example a simple greedy algorithm, where one takes an arbitrary ordering of the nodes and iteratively color each node with the first available color not already used in the neighbourhood. Obviously, this algorithm needs at most $p = \nu_1 + 1$ colors, where ν_1 is given by (3.35). Hence, the number of colors can be chosen independent of the level i as long as (3.35) is satisfied. Using the lexicographic ordering, the complexity of this algorithm is of order $\mathrm{O}(n_i)$ since the number of neighbours of each node is bounded. More sophisticated algorithms can decrease the number of colors even further, see for instance [PMX98] for a survey.

3.4.2. Dual norm estimates

We now consider the implementation of the multilevel preconditioners in Section 3.3.1 and the multilevel stationarity measure from Section 3.3.2. For this we have to calculate the scalar products (g_i, ϕ_k^j) for $k = 1, \dots, i$ and $j = 1, \dots, n_k$. For $k = i$ these are just the entries of $\underset{\sim}{g}_i$. Hence, we now suppose $k < i$. From the definition of the $\mathcal{U}$-orthogonal projection Q_k it follows that $(g_i, \phi_k^j) = (Q_k g_i, \phi_k^j)$. Since Q_k is also used as restriction in this setting, we analyze the complexity of this operation.

The spaces $\mathcal{V}_i$ are nested, which allows us to express each basis function $\phi_{i-1}^j \in \mathcal{V}_{i-1}$ in terms of the basis of $\mathcal{V}_i$, i.e., $\phi_{i-1}^j = \sum_{l=1}^{n_i} a_{lj} \phi_i^l$. Let us denote the matrix with the entries a_{lj} by $I_{i-1}^i \in \mathbb{R}^{n_i \times n_{i-1}}$. Then it holds:

$$(g_i, \phi_{i-1}^j) = (g_i, \sum_{l=1}^{n_i} a_{lj} \phi_i^l) = \sum_{l=1}^{n_i} a_{lj} (g_i, \phi_i^l) = \sum_{l=1}^{n_i} a_{lj} \underset{\sim}{g}_i^l.$$

This shows that the restriction of g_i represented by $\underset{\sim}{g}_i$ can be calculated by

$$\underset{\sim}{g}_{i-1} := \underset{\sim}{Q_{i-1} g_i} = (I_{i-1}^i)^T \underset{\sim}{g}_i.$$

In most applications the matrix I_{i-1}^i is sparse and cheap to assemble. For example in the case of a linear nodal basis that satisfies

$$\phi_i^j(x_i^l) = \delta_{jl} \quad \text{for all } x_i^l \in \mathcal{N}_i \text{ and } j = 1, \dots, n_i,$$

the entries match the values at the nodes, i.e., $a_{lj} = \phi_{i-1}^j(x_i^l)$.

A step $s_{i-1} \in \mathcal{V}_{i-1}$ is prolongated by means of the identity. Given a coefficient vector $\tilde{s}_{i-1} \in \mathbb{R}^{n_{i-1}}$, we seek the element in $\mathbb{R}^{n_i}$ that correspond to the same element in $\mathcal{V}_i$. From

$$s_{i-1} = \sum_{j=1}^{n_{i-1}} \tilde{s}_{i-1}^j \phi_{i-1}^j = \sum_{j=1}^{n_{i-1}} \Big(\tilde{s}_{i-1}^j \sum_{l=1}^{n_i} a_{lj} \phi_i^l \Big) = \sum_{l=1}^{n_i} \Big(\phi_i^l \sum_{j=1}^{n_{i-1}} \tilde{s}_{i-1}^j a_{lj} \Big).$$

we obtain that the entries of the coefficient vector are given by $\tilde{v}_i^l = \sum_{j=1}^{n_{i-1}} a_{lj} \tilde{s}_{i-1}^j$ and therefore $\tilde{v}_i = I_{i-1}^i \tilde{s}_{i-1}$ holds.

Remark 3.10 The matrices I_{i-1}^i and $(I_{i-1}^i)^T$ are similar to the restriction and interpolation operators in multigrid theory for finite differences. See also Example 2.2.

In the same way we can construct the matrices I_k^i for $k = 1, \dots, i-2$. Note that we also have the identity

$$I_k^i = I_k^{k+1} \cdots I_{i-1}^i \tag{3.59}$$

With these preliminaries we can formulate the preconditioners in terms of matrices and vectors. For ease of notation we set $I_i^i \in \mathbb{R}^{n_i \times n_i}$ to the identity matrix. The MDS preconditioner can be evaluated by

$$(g_i, \tilde{M}_i^{-1} g_i) = \sum_{k=1}^{i} \sum_{j=1}^{n_k} \frac{(g_i, \phi_k^j)^2}{(\phi_k^j, \phi_k^j)_{\mathcal{V}}} = \sum_{k=1}^{i} ((I_k^i)^T \underset{\sim}{g}_i)^T \operatorname{Diag}(\underset{\approx}{V}_k)^{-1} (I_k^i)^T \underset{\sim}{g}_i,$$

and for the BPX preconditioner we obtain

$$(g_i, \hat{M}_i^{-1} g_i) = \sum_{k=1}^{i} h_k^{2-d} \| (I_k^i)^T \underset{\sim}{g}_i \|_2^2.$$

We assume that there exists a $\delta < 1$, not depending on i or r, such that the number of unknowns satisfy $n_{k-1} \leq \delta n_k$ for $k = 1, \dots, r$. Typically, if we use uniform refinement, we get $\delta = 2^{-d}$. The evaluation of one summand in the preconditioner can be implemented using (3.59) with Cn_k operations and hence in total

$$ops = \sum_{k=1}^{i} Cn_k \leq C \sum_{k=1}^{i} \delta^{i-k} n_i = Cn_i \sum_{k=0}^{i-1} \delta^k \leq C \frac{1}{1-\delta} n_i.$$

So, independent of the number of levels, the costs for one evaluation is $\mathrm{O}(n_i)$.

4. Convexly constrained problems

In this chapter we consider problems whose feasible sets are convex. As in the preceding chapter, we will derive conditions under which smoothing steps yield a decrease of the quadratic model function that satisfies the fraction of Cauchy decrease condition (2.29) where the constant κ_{mdc} and β_C are level-independent. Furthermore, we will turn to a special class of constrained problems where the feasible set is a box. Here, we will show how to construct suitable lower-level boxes. Additionally, we will introduce an active-set strategy which changes the prolongation operators so that more directions are allowed in these lower-level boxes.

4.1. A level-independent stationarity measure

Considering our model setting from Example 3.1, we see that the projected gradient (2.26) as introduced in Chapter 2 is very expensive to evaluate. This has two main reasons: First, the gradient must be calculated with respect to the $H_0^1(\Omega)$ inner product on $\mathcal{V}$, which involves the solution of a PDE. Second, the projection on the feasible set must also be done with respect to the norm on $H_0^1(\Omega)$, which is even more expensive. What one would like to have is a stationarity measure where the gradient and the projection can be estimated cheaply at least for simple convex sets as for example pointwise bounds.

Similarly to the stationarity measure in the unconstrained case, one could try to use the dual norm of the projected gradient where both the representation as well as the projection is with respect to the $\mathcal{U}$-norm. It can be shown that such a measure is indeed a stationarity measure in the sense of Definition 2.2, but even in simple examples the continuity depends strongly on the fineness of the mesh, in comparison to the case without constraints. Since we are interested in level-independent quantities, we need a different measure.

4.1.1. A multilevel stationarity measure

In this section we introduce a new stationarity measure that uses the whole level hierarchy in the style of the measures introduced in Section 3.3.2.

We assume the variational setting from Section 3.1. Additionally, we introduce on each space $\mathcal{V}_i$ a level-dependent inner product $(\!(\cdot,\cdot)\!)_i$ and its associated norm $|||\cdot|||_i := \sqrt{(\!(\cdot,\cdot)\!)_i}$. We require the norm to be level-independently equivalent to the norm on $\mathcal{U}$, i.e., there exists a constant $C > 0$ such that

$$\frac{1}{C}\|u_i\| \leq |||u_i|||_i \leq C\|u_i\| \quad \text{for all } u_i \in \mathcal{V}_i,\ i = 1, \dots, r. \tag{4.1}$$

As in the previous chapter, we denote by C a generic constant, which is level-independent and is allowed to take different values in inequalities.

We define orthogonal projectors $Q_i^j \colon \mathcal{V}_i \to \mathcal{V}_j$, $j \leq i$, with respect to these inner products. This means they satisfy

$$(\!(Q_i^j v_i, v_j)\!)_j = (\!(v_i, v_j)\!)_i \quad \text{for all } v_j \in \mathcal{V}_j \text{ and } v_i \in \mathcal{V}_i. \tag{4.2}$$

In the first section no additional assumptions on the inner products are made. Hence, one is free to choose $(\!(\cdot,\cdot)\!)_i = (\cdot,\cdot)$. The main reason why we introduce these norms will become clear when we turn to box-constrained problems in Section 4.2.2.

The following assumption is fundamental for the multilevel stationarity measure we are going to introduce:

Assumption 4.1 *Let $g_i \in \mathcal{V}_i$. There exists a level-independent constant C such that*

$$C^{-1}\|\imath_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*}^2 \leq \sum_{j=1}^{i} (\lambda_j^{\max})^{-1}\|Q_j g_i\|^2 \leq C\|\imath_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*}^2 \quad \textit{for all } g_i \in \mathcal{V}_i,\ i = 1,\dots,r$$

is satisfied. As in the preceding chapter we denote by Q_j the $\mathcal{U}$-orthogonal projection onto $\mathcal{V}_j$.

Remark 4.1 Let $\{\lambda_j^{\max}\}_j$ satisfy the growth condition

$$\frac{\lambda_j^{\max}}{\lambda_{j-1}^{\max}} \geq \bar{\tau} > 1 \quad \text{for all } j = 2,\dots,r.$$

Then, in the setting of Example 3.1, Assumption 4.1 follows from the famous equivalence (cf. for instance [Osw94, Yse93, BY93])

$$C^{-1}\|v_i\|_{\mathcal{V}_i}^2 \leq \lambda_1^{\max}\|Q_1 v_i\|^2 + \sum_{j=1}^{i} \lambda_j^{\max}\|(Q_j - Q_{j-1})v_i\|^2 \leq C\|v_i\|_{\mathcal{V}_i}^2 \text{ for all } v_i \in \mathcal{V}_i \tag{4.3}$$

by duality arguments. To verify this, first note that we can rewrite $v_i \in \mathcal{V}_i$ as

$$v_i = Q_1 v_i + \sum_{j=2}^{i} (Q_j - Q_{j-1})v_i,$$

and since $(Q_j - Q_{j-1})v_i$ is orthogonal on $\mathcal{V}_{j-1}$, we have

$$(v_i, w_i) = (Q_1 v_i, Q_1 w_i) + \sum_{j=2}^{i} ((Q_j - Q_{j-1})v_i, (Q_j - Q_{j-1})w_i).$$

Using the Cauchy-Schwarz inequality twice, we obtain

$$\begin{aligned}
(v_i, w_i) &\leq (\lambda_1^{\max})^{-1/2}\|Q_1 v_i\|(\lambda_1^{\max})^{1/2}\|Q_1 w_i\| \\
&\quad + \sum_{j=2}^{i} (\lambda_j^{\max})^{-1/2}\|(Q_j - Q_{j-1})v_i\|(\lambda_j^{\max})^{1/2}\|(Q_j - Q_{j-1})w_i\| \\
&\leq \Big((\lambda_1^{\max})^{-1}\|Q_1 v_i\|^2 + \sum_{j=2}^{i} (\lambda_j^{\max})^{-1}\|(Q_j - Q_{j-1})v_i\|^2\Big)^{1/2} \\
&\quad \cdot \Big(\lambda_1^{\max}\|Q_1 w_i\|^2 + \sum_{j=2}^{i} \lambda_j^{\max}\|(Q_j - Q_{j-1})w_i\|^2\Big)^{1/2}.
\end{aligned}$$

By definition of the dual norm, (4.3) and the last estimate it follows

$$\begin{aligned}\|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*} &= \sup_{0\neq w_i\in\mathcal{V}_i}\frac{(g_i,w_i)}{\|w_i\|_{\mathcal{V}_i}} \leq C\sup_{0\neq w_i\in\mathcal{V}_i}\frac{(g_i,w_i)}{\Big(\lambda_1^{\max}\|Q_1w_i\|^2+\sum_{j=1}^i\lambda_j^{\max}\|(Q_j-Q_{j-1})w_i\|^2\Big)^{1/2}}\\ &\leq C\Big((\lambda_1^{\max})^{-1}\|Q_1g_i\|^2+\sum_{j=2}^i(\lambda_j^{\max})^{-1}\|(Q_j-Q_{j-1})g_i\|^2\Big)^{1/2}.\end{aligned}$$

To show the other direction, we set

$$\bar{w}_i=(\lambda_1^{\max})^{-1}Q_1g_i+\sum_{j=2}^i(\lambda_j^{\max})^{-1}(Q_j-Q_{j-1})g_i.$$

Note that $(Q_k-Q_{k-1})\bar{w}_i=(\lambda_k^{\max})^{-1}(Q_k-Q_{k-1})g_i$ and $Q_1\bar{w}_i=(\lambda_1^{\max})^{-1}Q_1g_i$ holds. Furthermore,

$$(\bar{w}_i,g_i)=(\lambda_1^{\max})^{-1}\|Q_1g_i\|^2+\sum_{j=2}^i(\lambda_j^{\max})^{-1}\|(Q_j-Q_{j-1})g_i\|^2.$$

Using this special element and (4.3), one obtains the other inequality:

$$\begin{aligned}\|\iota_{\mathcal{U}}(g_i)\|_{\mathcal{V}_i^*} &\geq \frac{(g_i,\bar{w}_i)}{\|\bar{w}_i\|_{\mathcal{V}_i}} \geq \frac{1}{C}\frac{(g_i,\bar{w}_i)}{\Big(\lambda_1^{\max}\|Q_1\bar{w}_i\|^2+\sum_{j=1}^i\lambda_j^{\max}\|(Q_j-Q_{j-1})\bar{w}_i\|^2\Big)^{1/2}}\\ &= \frac{1}{C}\Big((\lambda_1^{\max})^{-1}\|Q_1g_i\|^2+\sum_{j=2}^i(\lambda_j^{\max})^{-1}\|(Q_j-Q_{j-1})g_i\|^2\Big)^{1/2}.\end{aligned}$$

We finish this remark by noting that

$$\begin{aligned}(\lambda_1^{\max})^{-1}\|Q_1g_i\|^2+\sum_{j=2}^i(\lambda_j^{\max})^{-1}\|(Q_j-Q_{j-1})g_i\|^2 &= \sum_{j=1}^{i-1}((\lambda_j^{\max})^{-1}-(\lambda_{j+1}^{\max})^{-1})\|Q_jg_i\|^2\\ &\quad+(\lambda_i^{\max})^{-1}\|Q_ig_i\|^2\\ &\geq(1-\bar{\tau}^{-1})\sum_{j=1}^i(\lambda_j^{\max})^{-1}\|Q_jg_i\|^2\end{aligned}$$

and

$$\sum_{j=1}^{i-1}((\lambda_j^{\max})^{-1}-(\lambda_{j+1}^{\max})^{-1})\|Q_jg_i\|^2+(\lambda_i^{\max})^{-1}\|Q_ig_i\|^2\leq\sum_{j=1}^i(\lambda_j^{\max})^{-1}\|Q_jg_i\|^2$$

holds.

Given the problem

$$\min_{v_i\in C_i}h_i(v_i), \tag{4.4}$$

with a closed, convex and nonempty set $C_i \subset \mathcal{V}_i$ we define the *multilevel stationarity measure* $\chi_i^{\mathrm{ML}} \colon C_i \to \mathbb{R}_+$ to (4.4) by

$$\chi_i^{\mathrm{ML}}(v_i) := \left(\sum_{j=1}^{i} (\lambda_j^{\max})^{-1} |||\mathrm{Proj}_{C_j(v_i)}(-Q_i^j \nabla h_i(v_i))|||_j^2 \right)^{1/2}, \tag{4.5}$$

where $C_i(v_i) := C_i - v_i$, and $C_j(v_i) \subset \mathcal{V}_j$ are closed convex sets that satisfy

$$0 \in C_j(v_i) \quad \text{and} \quad s_j \in C_j(v_i) \Rightarrow v_i + s_j \in C_i.$$

The gradient $\nabla h_i(v_i) \in \mathcal{V}_i$ denotes here the representation of $h_i'(v_i)$ with respect to $(\!(\cdot,\cdot)\!)_i$, i.e.,

$$\langle h_i'(v_i), u_i \rangle = (\!(\nabla h_i(v_i), u_i)\!)_i \quad \text{for all } u_i \in \mathcal{V}_i.$$

The operator $\mathrm{Proj}_{C_j(v_i)}$ is assumed to be the orthogonal projection onto $C_j(v_i)$ with respect to $(\!(\cdot,\cdot)\!)_j$ for all j.

We start our analysis of χ_i^{ML} by showing that it satisfies (2.25).

Lemma 4.1 *The function χ_i^{ML} satisfies*

$$\chi_i^{ML}(v_i^*) = 0 \text{ if and only if } v_i^* \text{ is a KKT-Point of } \min_{v_i \in C_i} h_i(v_i).$$

Proof We first show the following implication:

$$|||\mathrm{Proj}_{C_i(v_i^*)}(-\nabla h_i(v_i^*))|||_i = 0 \quad \Rightarrow \quad |||\mathrm{Proj}_{C_j(v_i^*)}(-Q_i^j \nabla h_i(v_i^*))|||_j = 0 \text{ for } j = 1, \dots, i-1. \tag{4.6}$$

By definition of the sets $C_j(v_i^*)$, $C_j(v_i^*) \subset C_i(v_i^*)$ holds. If $\mathrm{Proj}_{C_i(v_i^*)}(-\nabla h_i(v_i^*))) = 0$, it follows from the Projection Theorem A.2 that

$$(\!(\nabla h_i(v_i^*), v_i)\!)_i \geq 0 \quad \text{for all } v_i \in C_i(v_i^*),$$

and hence for $j = 1, \dots, i-1$

$$0 \leq (\!(\nabla h_i(v_i^*), v_j)\!)_i = (\!(Q_i^j \nabla h_i(v_i^*), v_j)\!)_j \quad \text{for all } v_j \in C_j(v_i^*)$$

is satisfied. Another application of the Projection Theorem yields assertion (4.6).

After this prerequisite, it remains to show that $\mathrm{Proj}_{C_i(v_i^*)}(-\nabla h_i(v_i^*)) = 0$ iff v_i^* is a KKT-Point. Let v_i^* be a KKT-Point, i.e., it satisfies

$$0 \leq \langle h_i'(v_i^*), v_i - v_i^* \rangle = (\!(\nabla h_i(v_i^*), v_i - v_i^*)\!)_i \text{ for all } v_i \in C_i.$$

Using the set $C_i(v_i^*)$, this can be written as

$$0 \leq (\!(\nabla h_i(v_i^*), v_i)\!)_i \quad \text{for all } v_i \in C_i(v_i^*).$$

According to the Projection Theorem, this is equivalent to $\mathrm{Proj}_{C_i(v_i^*)}(-\nabla h_i(v_i^*)) = 0$. □

If the problem is unconstrained, the measure is equivalent to the dual norm of the derivative. This is the assertion of the next lemma.

Lemma 4.2 *Let Assumption 4.1 hold. Then*

$$\frac{1}{C}\|h_i'(v_i)\|_{\mathcal{V}_i^*}^2 \leq \sum_{j=1}^{i}(\lambda_j^{\max})^{-1}|||Q_i^j\nabla h_i(v_i)|||_j^2 \leq C\|h_i'(v_i)\|_{\mathcal{V}_i^*}^2 \quad \text{for all } v_i \in \mathcal{V}_i$$

is satisfied. Moreover, if $C_i = \mathcal{V}_i$ and the lower-level feasible sets are chosen as $C_j(v_i) = \mathcal{V}_j$, then there holds

$$\frac{1}{C}\|h_i'(v_i)\|_{\mathcal{V}_i^*} \leq \chi_i^{ML}(v_i) \leq C\|h_i'(v_i)\|_{\mathcal{V}_i^*} \quad \text{for all } v_i \in \mathcal{V}_i.$$

Proof Assume that g_i and $\bar{g}_i$ are elements of $\mathcal{V}_i$ that satisfy

$$(g_i, v_i) = (\!(\bar{g}_i, v_i)\!)_i \text{ for all } v_i \in \mathcal{V}_i.$$

Using the definitions of Q_j and Q_i^j we have for all $v_j \in \mathcal{V}_j$:

$$(Q_j g_i, v_j) = (g_i, v_j) = (\!(\bar{g}_i, v_j)\!)_i = (\!(Q_i^j\bar{g}_i, v_j)\!)_j.$$

The equivalence of the norms $|||\cdot|||_j$ and $\|\cdot\|$ on $\mathcal{V}_j$ yields

$$\|Q_j g_i\| = \max_{0\neq v_j\in\mathcal{V}_j}\frac{(Q_j g_i, v_j)}{\|v_j\|} = \max_{0\neq v_j\in\mathcal{V}_j}\frac{(\!(Q_i^j\bar{g}_i, v_j)\!)_j}{\|v_j\|} \leq C\max_{0\neq v_j\in\mathcal{V}_j}\frac{(\!(Q_i^j\bar{g}_i, v_j)\!)_j}{|||v_j|||_j} = C|||Q_i^j\bar{g}_i|||_j.$$

In the same way, one shows $|||Q_i^j\bar{g}_i|||_j \leq C\|Q_j g_i\|$. Hence, if we replace $|||Q_i^j\nabla h_i(v_i)|||_j$ by $\|Q_i\nabla_{\mathcal{U}}h_i(v_i)\|$, where $\nabla_{\mathcal{U}}h_i(v_i)$ is the $\mathcal{U}$-representation of $h_i'(v_i)$, we obtain an equivalent stationarity measure. Using Assumption 4.1 now shows the assertion.

The second assertion follows directly from the fact that under the stated assumptions the stationarity measure becomes

$$\chi_i^{\mathrm{ML}}(v_i) = \Big(\sum_{j=1}^{i}(\lambda_j^{\max})^{-1}|||Q_i^j\nabla h_i(v_i)|||_j^2\Big)^{1/2}.$$

□

4.1.2. Continuity of χ_i^{ML}

In order to analyse the continuity of the new stationarity measure χ_i^{ML}, we need a concept of continuity for set-valued mappings.

Definition 4.1 Let M be a normed space. The *Hausdorff distance* $d_H\colon \mathcal{P}(M)\times\mathcal{P}(M) \to [0,\infty]$[1] of two sets $A, B \subset M$ is defined by

$$d_H(A,B) := \max\{\sup_{v\in A} d(v,B), \sup_{v\in B} d(v,A)\},$$

[1] By $\mathcal{P}(M) := \{A \mid A \subset M\}$ we denote the powerset of M.

where the distance d is given by

$$d(v, A) := \inf_{w \in A} \|v - w\|.$$

A sequence of sets $(A_n)_{n\in\mathbb{N}}$, $A_n \subset M$ converges to a set A in the Hausdorff sense iff

$$\lim_{n\to\infty} d_H(A_n, A) = 0.$$

The next lemma gives an estimate for the distance between the projections on different sets.

Lemma 4.3 *Let H be a Hilbert space and A and B be closed convex subsets of H that both contain 0. Then it holds for all $x, y \in H$:*

$$\|Proj_A(x) - Proj_B(y)\| \leq \sqrt{2\min\{\|x\|, \|y\|\} d_H(A, B)} + \|x - y\|,$$

where $Proj_A$ ($Proj_B$) denotes the H-orthogonal projection onto A (B).

Proof We recall that the projection on a closed convex set is well-defined and unique, since H is a Hilbert space (cf. Theorem A.2). We first derive an estimate for the simpler case $x = y$. We have

$$\begin{aligned}\|\mathrm{Proj}_A(x) - \mathrm{Proj}_B(x)\|^2 &= (\mathrm{Proj}_A(x) - \mathrm{Proj}_B(x), \mathrm{Proj}_A(x) - \mathrm{Proj}_B(x)) \\ &= (\mathrm{Proj}_A(x) - x, \mathrm{Proj}_A(x) - \mathrm{Proj}_B(x)) \\ &\quad + (\mathrm{Proj}_B(x) - x, \mathrm{Proj}_B(x) - \mathrm{Proj}_A(x)).\end{aligned}$$

From the definition of the Hausdorff distance follows the existence of an element $z_1 \in A$ with $\|z_1 - \mathrm{Proj}_B(x)\| \leq d_H(A, B)$. Using z_1, we estimate

$$\begin{aligned}(\mathrm{Proj}_A(x) - x, \mathrm{Proj}_A(x) - \mathrm{Proj}_B(x)) &= (\mathrm{Proj}_A(x) - x, \mathrm{Proj}_A(x) - z_1) \\ &\quad + (\mathrm{Proj}_A(x) - x, z_1 - \mathrm{Proj}_B(x)) \\ &\leq \|\mathrm{Proj}_A(x) - x\| d_H(A, B),\end{aligned}$$

where we have used that the first term is negative, which follows from the Projection Theorem A.2. Since $0 \in A$, it follows further that

$$\|\mathrm{Proj}_A(x) - x\| = \min_{y\in A} \|y - x\| \leq \|0 - x\|$$

holds and hence

$$(\mathrm{Proj}_A(x) - x, \mathrm{Proj}_A(x) - \mathrm{Proj}_B(x)) \leq \|x\| d_H(A, B).$$

In the same way, we obtain

$$(\mathrm{Proj}_B(x) - x, \mathrm{Proj}_B(x) - \mathrm{Proj}_A(x)) \leq \|x\| d_H(A, B).$$

Hence, we have

$$\|\mathrm{Proj}_A(x) - \mathrm{Proj}_B(x)\|^2 \leq 2\|x\| d_H(A, B). \tag{4.7}$$

In the case $x \neq y$ we use triangle inequality and the Lipschitz continuity of the projection (cf. Lemma A.1):

$$\|\mathrm{Proj}_A(x) - \mathrm{Proj}_B(y)\| \leq \|\mathrm{Proj}_A(x) - \mathrm{Proj}_B(x)\| + \|x - y\| \leq \sqrt{2\|x\| d_H(A, B)} + \|x - y\|.$$

The observation that we can switch the roles of x and y finishes the proof. □

Remark 4.2 A similar estimate was proved in [AN95] for the more general case of projections in Banach spaces, which is a lot more technical compared to our setting.

The following simple lemma is needed for our main theorem:

Lemma 4.4 *Let $(a_k)_{k=1}^n$ and $(b_k)_{k=1}^n$ be sequences with elements a_k, b_k belonging to a Banach space with norm $\|\cdot\|$. The following estimate holds:*

$$\Big|\sum_{k=1}^n \big(\|a_k\|^2 - \|b_k\|^2\big)\Big| \le \sum_{k=1}^n \|a_k - b_k\|^2 + 2\Big(\sum_{k=1}^n \|b_k\|^2\Big)^{1/2}\Big(\sum_{k=1}^n \|a_k - b_k\|^2\Big)^{1/2}.$$

Proof The assertion follows easily with the inverse triangle and the Cauchy-Schwarz inequality:

$$\begin{aligned}\Big|\sum_{k=1}^n \big(\|a_k\|^2 - \|b_k\|^2\big)\Big| &\le \sum_{k=1}^n \big|\big(\|a_k\| - \|b_k\|\big)^2 + 2\|b_k\|\big(\|a_k\| - \|b_k\|\big)\big| \\ &\le \sum_{k=1}^n \|a_k - b_k\|^2 + 2\Big(\sum_{k=1}^n \|b_k\|^2\Big)^{1/2}\Big(\sum_{k=1}^n \big(\|a_k\| - \|b_k\|\big)^2\Big)^{1/2} \\ &\le \sum_{k=1}^n \|a_k - b_k\|^2 + 2\Big(\sum_{k=1}^n \|b_k\|^2\Big)^{1/2}\Big(\sum_{k=1}^n \|a_k - b_k\|^2\Big)^{1/2}. \qquad \square\end{aligned}$$

We now show the continuity of χ_i^{ML} under suitable assumptions. Since we are interested in the continuity with respect to the level i, we explicitly estimate the size of the δ in the ε-δ definition of continuity. We will later use these estimates to make a more extensive analysis in the special case of box-constrained problems.

Theorem 4.1 *Assume that $h_i' \colon C_i \to \mathcal{V}_i^*$ is continuous, i.e., for every $\varepsilon_g > 0$ and every $v_i \in C_i$ exists a $\delta_g(v_i, \varepsilon_g) > 0$ such that*

$$\|h_i'(v_i) - h_i'(u_i)\|_{\mathcal{V}_i^*} \le \varepsilon_g \text{ for all } u_i \in C_i \text{ with } \|v_i - u_i\|_{\mathcal{V}_i} \le \delta_g(v_i, \varepsilon_g). \tag{4.8}$$

Furthermore, suppose that

$$d_H(C_j(v_i), C_j(u_i)) \le C c_j \|v_i - u_i\|_{\mathcal{V}_i} \text{ for all } j = 1, \dots, i-1, \tag{4.9}$$

where c_j is a constant which depends on j. Then χ_i^{ML} is continuous on C_i, more precisely for every $\varepsilon > 0$ and every $v_i \in C_i$ it holds:

$$|\chi_i^{ML}(v_i) - \chi_i^{ML}(u_i)| \le \varepsilon \text{ for all } u_i \in C_i \text{ with } \|v_i - u_i\|_{\mathcal{V}_i} \le \delta(v_i, \varepsilon), \tag{4.10}$$

where $\delta(v_i, \varepsilon) := \min\{\varepsilon_g^2,\ \delta_g(v_i, \varepsilon_g)\}$ with

$$\varepsilon_g := \min\left\{1,\ \frac{\varepsilon^2}{C \max\left\{1, \|h_i'(v_i)\|_{\mathcal{V}_i^*}^{3/2}\right\}(B_i + 1)}\right\} \tag{4.11}$$

and

$$B_i^2 := \max\Big\{1, \sum_{j=1}^i (\lambda_j^{\max})^{-1} c_j^2\Big\}. \tag{4.12}$$

PROOF For brevity we set $g_j(v_i) = Q_i^j \nabla h_i(v_i)$. We start by estimating the difference $|||\mathrm{Proj}_{C_j(v_i)}(-g_j(v_i)) - \mathrm{Proj}_{C_j(u_i)}(-g_j(u_i))|||_j^2$ using Lemma 4.3 and (4.9):

$$\begin{aligned}|||\mathrm{Proj}_{C_j(v_i)}(-g_j(v_i)) - \mathrm{Proj}_{C_j(u_i)}(-g_j(u_i))|||_j^2 &\le \Big(\sqrt{2|||g_j(v_i)|||_j d_H(C_j(v_i), C_j(u_i))} \\ &\quad + |||g_j(v_i) - g_j(u_i)|||_j\Big)^2 \\ &\le 4|||g_j(v_i)|||_j d_H(C_j(v_i), C_j(u_i)) + 2|||g_j(v_i) - g_j(u_i)|||_j^2 \\ &\le Cc_j|||g_j(v_i)|||_j \|u_i - v_i\|_{\mathcal{V}_i} + 2|||g_j(v_i) - g_j(u_i)|||_j^2.\end{aligned}$$

Although pessimistic, this estimate is also true for $j = i$. Summing over all levels and using the Cauchy-Schwarz inequality yields

$$\begin{aligned}A^2 &:= \sum_{j=1}^{i} (\lambda_j^{\max})^{-1} |||\mathrm{Proj}_{C_j(v_i)}(-g_j(v_i)) - \mathrm{Proj}_{C_j(u_i)}(-g_j(u_i))|||_j^2 \\ &\le C\|u_i - v_i\|_{\mathcal{V}_i} \Big(\sum_{j=1}^{i} (\lambda_j^{\max})^{-1} |||g_j(v_i)|||_j^2\Big)^{1/2} \Big(\sum_{j=1}^{i} (\lambda_j^{\max})^{-1} c_j^2\Big)^{1/2} \\ &\quad + 2\sum_{j=1}^{i} (\lambda_j^{\max})^{-1} |||g_j(v_i) - g_j(u_i)|||_j^2 \\ &\le C\|u_i - v_i\|_{\mathcal{V}_i} \|h_i'(v_i)\|_{\mathcal{V}_i^*} B_i + 2C\|h_i'(v_i) - h_i'(u_i)\|_{\mathcal{V}_i^*}^2.\end{aligned}$$

Note that we have used Lemma 4.2 in the last step.

Let $0 < \varepsilon_g \le 1$ be arbitrary. From the continuity of h_i' follows

$$A^2 \le C(\|h_i'(v_i)\|_{\mathcal{V}_i^*} \delta_g(v_i, \varepsilon_g) B_i + \varepsilon_g^2) \text{ for all } u_i \text{ with } \|u_i - v_i\|_{\mathcal{V}_i} \le \delta_g(v_i, \varepsilon_g).$$

In the following, we assume that $\|u_i - v_i\|_{\mathcal{V}_i} \le \delta_g(v_i, \varepsilon_g)$ holds. Since χ_i^{ML} is non-negative, we have

$$\begin{aligned}|\chi_i^{\mathrm{ML}}(v_i) - \chi_i^{\mathrm{ML}}(u_i)|^2 &\le |\chi_i^{\mathrm{ML}}(v_i)^2 - \chi_i^{\mathrm{ML}}(u_i)^2| \\ &= \Big|\sum_{j=1}^{i} (\lambda_j^{\max})^{-1} (|||\mathrm{Proj}_{C_j(v_i)}(-g_j(v_i))|||_j^2 - |||\mathrm{Proj}_{C_j(u_i)}(-g_j(u_i))|||_j^2)\Big|.\end{aligned}$$

From the previous technical lemma we infer

$$|\chi_i^{\mathrm{ML}}(v_i) - \chi_i^{\mathrm{ML}}(u_i)|^2 \le A^2 + 2\Big(\sum_{j=1}^{i} (\lambda_j^{\max})^{-1} |||\mathrm{Proj}_{C_j(v_i)}(-g_j(v_i))|||_j^2\Big)^{1/2} A.$$

Inserting our estimate of A in the last expression and using that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$, yields

$$\begin{aligned}|\chi_i^{\mathrm{ML}}(v_i) - \chi_i^{\mathrm{ML}}(u_i)|^2 &\le C(\|h_i'(v_i)\|_{\mathcal{V}_i^*} \delta_g(v_i, \varepsilon_g) B_i + \varepsilon_g^2) \\ &\quad + C\chi_i^{\mathrm{ML}}(v_i) \left(\|h_i'(v_i)\|_{\mathcal{V}_i^*}^{1/2} \delta_g(v_i, \varepsilon_g)^{1/2} B_i^{1/2} + \varepsilon_g\right).\end{aligned}$$

Since $0 \in C_j(v_i)$ and the projection is Lipschitz continuous with Lipschitz constant one, we can estimate

$$\begin{aligned}\chi_i^{\mathrm{ML}}(v_i)^2 &= \sum_{j=1}^{i} (\lambda_j^{\max})^{-1} \left(|||\mathrm{Proj}_{C_j(v_i)}(-g_j(v_i))|||_j - |||\mathrm{Proj}_{C_j(v_i)}(0)|||_j \right)^2 \\ &\leq \sum_{j=1}^{i} (\lambda_j^{\max})^{-1} |||g_j(v_i)|||_j^2 \leq C \|h_i'(v_i)\|_{\mathcal{V}_i^*}^2.\end{aligned} \tag{4.13}$$

Without loss of generality, we assume that $\delta_g(v_i, \varepsilon_g) \leq \varepsilon_g^2$ and since $\varepsilon_g \leq 1$ we also have $\varepsilon_g^2 \leq \varepsilon_g$. Hence, using (4.13) we obtain

$$|\chi_i^{\mathrm{ML}}(v_i) - \chi_i^{\mathrm{ML}}(u_i)|^2 \leq C\varepsilon_g \max\{1, \|h_i'(v_i)\|_{\mathcal{V}_i^*}^{3/2}\}(B_i + 1).$$

Inserting (4.11) into the last statement with a suitable constant C, yields $|\chi_i^{\mathrm{ML}}(v_i) - \chi_i^{\mathrm{ML}}(u_i)| \leq \varepsilon$ for all u_i with $\|u_i - v_i\|_{\mathcal{V}_i} \leq \delta(v_i, \varepsilon) := \min\{\varepsilon_g^2, \delta_g(v_i, \varepsilon_g)\}$. This finishes the proof. □

Under the assumptions of the last theorem, it follows with Lemma 4.1 that χ_i^{ML} is a stationarity measure according to Definition 2.2.

We finally analyze in which cases we have the stronger property that χ_i^{ML} is uniformly continuous, which is needed for the strong convergence result in Theorem 2.2. This is the assertion of the following corollary:

Corollary 4.1 *Let h_i' be uniformly continuous on a set $\mathcal{S}_i \subset \mathcal{V}_i$, i.e., there exists $\delta_g(\varepsilon_g) > 0$ independent of v_i such that (4.8) holds with $\delta_g(v_i, \varepsilon_g) \geq \delta_g(\varepsilon_g)$ for all $v_i \in \mathcal{S}_i$. If furthermore $\|h_i'(v_i)\|_{\mathcal{V}_i^*}$ is bounded by a constant β_i for all $v_i \in \mathcal{S}_i$, then χ_i^{ML} is uniformly continuous on $\mathcal{S}_i$.*

Proof If $\|h_i'(v_i)\|_{\mathcal{V}_i^*}$ is bounded, the choice of ε_g in the previous theorem can be done independent of v_i. From the uniform continuity of h_i' finally follows then that we can choose $\delta(\varepsilon) \leq \min\{\varepsilon_g^2, \delta_g(\varepsilon_g)\}$ independent of $v_i \in \mathcal{S}_i$ such that (4.10) is satisfied with $\delta(\varepsilon, v_i) \geq \delta(\varepsilon)$. □

We are also interested, whether the uniform continuity is level-independent, meaning that the choice of $\delta(\varepsilon)$ does not depend on the mesh-size h_i or the number of levels used. In a typical setting where the underlying infinite dimensional functional is uniform continuous, the answer to this question depends mainly on the construction of the lower-level sets. We will analyse the level dependency for our typical setting in Section 4.3.1.

4.2. Level independent Cauchy decrease

For a given iterate $v_{i,k}$, we consider the quadratic trust-region subproblem

$$\begin{aligned}&\min_{s_i \in \mathcal{V}_i} q_i(s_i) := (\!(s_i,\, g_i)\!)_i + \frac{1}{2}(\!(s_i,\, H_i s_i)\!)_i \\ &\text{subject to } \|s_i\|_i \leq \Delta_i, \quad s_i \in C_i.\end{aligned} \tag{4.14}$$

Here, $g_i \in \mathcal{V}_i$ is the gradient of h_i in terms of the new inner product, i.e., $(\!(s_i, g_i)\!)_i = \langle h_i'(v_{i,k}), s_i \rangle$ holds for all $s_i \in \mathcal{V}_i$. The linear operator $H_i \colon \mathcal{V}_i \to \mathcal{V}_i$ is an approximation of the second derivative of $h_i(v_{i,k})$. As in the unconstrained case we demand that H_i is symmetric and satisfies (3.23), i.e.,

$$(\!(u_i, H_i v_i)\!)_i \leq C_H \lambda_i^{\max} \|u_i\| \, \|v_i\| \quad \text{for all } u_i, v_i \in \mathcal{V}_i.$$

From the norm equivalence follows

$$(\!(u_i, H_i v_i)\!)_i \leq C C_H \lambda_i^{\max} |\!|\!| u_i |\!|\!|_i |\!|\!| v_i |\!|\!|_i. \tag{4.15}$$

Note that in contrast to problem (2.32), we assume without loss of generality $v_{i,k} = 0$ in the constraints. This can be achieved by replacing C_i in (2.32) by the convex set

$$C_i(v_{i,k}) := \{ s_i \in \mathcal{V}_i \,|\, v_{i,k} + s_i \in C_i \}.$$

We always assume $0 \in C_i$.

For the rest of this section, we will consider algorithms that approximately solve the trust-region problem (4.14). To show a level-independent Cauchy-decrease, we demand an assumption similar to Assumption 3.1 in the unconstrained case:

Assumption 4.2 *If $v_{i,k} \in \mathcal{V}_i$ violates the smoothness property* (2.30)*, i.e.*

$$\chi_{i-1}(0) < \kappa_\chi \chi_i(v_{i,k}) \tag{4.16}$$

holds, then

$$|\!|\!| p_i |\!|\!|_i^2 \geq c(\kappa_\chi, \tau)^2 \lambda_i^{\max} \chi_i(v_{i,k})^2$$

is satisfied, where $p_i = Proj_{C_i(v_i)}(-\nabla h_i(v_{i,k}))$ is the projected gradient with respect to $(\!(\cdot,\cdot)\!)_i$. $c(\kappa_\chi, \tau) > 0$ denotes a level-independent constant that depends on κ_χ and τ.

The proof of the following lemma is omitted, since it follows directly from the definition of χ_i^{ML}.

Lemma 4.5 *The stationarity measure χ_i^{ML} satisfies Assumption 4.2 with $c(\kappa_\chi, \tau) = \sqrt{1 - \kappa_\chi^2}$.*

Since we consider a trust-region subproblem in a fixed iteration k, we omit this index in the following.

4.2.1. A projected gradient step

Similarly to the unconstrained case, we first consider the simple step $s_i^* = t^* p_i$ where t^* is the solution of the one dimensional problem

$$\begin{aligned} &\min_{t \in \mathbb{R}_+} \psi_i(t) := q_i(t p_i) = t (\!(p_i,\, g_i)\!)_i + \frac{t^2}{2} (\!(p_i,\, H_i p_i)\!)_i \\ &\text{subject to } t \leq \frac{\Delta_i}{\|p_i\|_i}, \quad t p_i \in C_i, \end{aligned} \tag{4.17}$$

and

$$p_i := \mathrm{Proj}_{C_i}(-g_i).$$

If p_i is not smooth, the next lemma shows that the fraction of Cauchy decrease condition (2.29) holds for the step s_i^* with constants that are level-independent.

Lemma 4.6 *Let χ_i satisfy Assumption 4.2. Suppose that t^* is the solution of* (4.17) *and let the trust-region norms satisfy*

$$\|v_i\|_i \leq C\lambda_i^{\max}|||v_i|||_i \quad \textit{for all } v_i \in \mathcal{V}_i. \tag{4.18}$$

If (4.16) *holds, the step $s_i^* := t^* p_i$ satisfies* (2.29)*; more precisely the predicted reduction can be estimated by*

$$-q_i(s_i^*) \geq C^{-1} c(\kappa_\chi, \tau)\chi_i(v_i) \min\left\{\Delta_i, \frac{c(\kappa_\chi,\tau)}{C_H}\chi_i(v_i)\right\},$$

with a level-independent constant C.

Proof We first analyse the case $(\!(p_i,\, H_i p_i)\!)_i > 0$. A simple calculation shows that

$$\hat{t} := -\frac{(\!(p_i, g_i)\!)_i}{(\!(p_i, H_i p_i)\!)_i}$$

is the global minimum of ψ_i. From (4.15) we obtain

$$\psi_i(\hat{t}) = -\frac{1}{2}\frac{(\!(p_i, g_i)\!)_i^2}{(\!(p_i, H_i p_i)\!)_i} \leq -\frac{1}{CC_H\lambda_i^{\max}}\frac{(\!(p_i, g_i)\!)_i^2}{|||p_i|||_i^2}.$$

From the Projection Theorem (A.2), it follows

$$|||p_i|||_i^2 = (\!(-g_i, p_i)\!)_i - \underbrace{(\!(-g_i - p_i, p_i)\!)_i}_{\geq 0} \leq (\!(-g_i, p_i)\!)_i \tag{4.19}$$

and thus $-\psi_i(\hat{t}) \geq \frac{1}{C\tilde{C}\lambda_i^{\max}}|||p_i|||_i^2$. Let us now assume that $t^* < \hat{t}$, which is the case when the step lies on the boundary of the feasible set. Then from the definition of $\hat{t}$ follows $-(\!(p_i, g_i)\!)_i > t^*(\!(p_i, H_i, p_i)\!)_i$ and hence with (4.19):

$$\psi_i(t^*) = t^*(\!(p_i, g_i)\!)_i + \frac{t^*}{2}(\!(p_i, H_i, p_i)\!)_i < \frac{t^*}{2}(\!(p_i, g_i)\!)_i \leq -\frac{t^*}{2}|||p_i|||_i^2.$$

When the stepsize is limited by the trust-region condition, i.e., $t^* = \frac{\Delta_i}{\|p_i\|_i}$, it follows from (4.18) that

$$-\psi_i(t^*) \geq \frac{\Delta_i}{2}\frac{|||p_i|||_i^2}{\|p_i\|_i} \geq \frac{\Delta_i}{C\sqrt{\lambda_i^{\max}}}|||p_i|||_i$$

holds. Otherwise, if the step length t^* is limited by the convex set, i.e., $t^* = \max\{t > 0 \,|\, tp_i \in C_i\}$, we infer from the definition of p_i that $t^* \geq 1$ and hence

$$-\psi_i(t^*) \geq \frac{1}{2}|||p_i|||_i^2.$$

We recall that by definition $\lambda_i^{\max} \geq 1$ and therefore, in either case, we can estimate the descent of $t^* p_i$ by

$$-\psi_i(t^*) = -q_i(t^* p_i) \geq \frac{1}{C\sqrt{\lambda_i^{\max}}} |||p_i|||_i \min\left\{\Delta_i, \frac{|||p_i|||_i}{C_H \sqrt{\lambda_i^{\max}}}\right\}. \tag{4.20}$$

If the curvature of H_i in direction p_i is not positive, i.e., $(\!(p_i, H_i p_i)\!)_i \leq 0$, ψ_i is unbounded for $t \to \infty$ and therefore the minimum lies at the boundary of the feasible set. As above we have $t^* \geq \min\{\Delta_i / \|p_i\|_i, 1\}$. This leads to

$$-\psi_i(t^*) \geq -t^*(\!(p_i, g_i)\!)_i \geq \frac{1}{C} \min\left\{\frac{\Delta_i}{\sqrt{\lambda_i^{\max}}}, |||p_i|||_i\right\} |||p_i|||_i.$$

Thus the step also satisfies (4.20) in this case. Now the final descent estimate follows from (4.16) and by applying Assumption 4.2 to (4.20). □

Instead of the step s_i^* from the preceding lemma, another common choice is an approximate minimizer of the trust-region subproblem along the projected gradient path $s_i(t) := \mathrm{Proj}_{C_i}(-t g_i)$, $t > 0$. Such algorithms were studied for instance in [Toi88] or [CGT00, Sec. 12.2]. We will outline a method presented in [Mor88].

Let $0 < \mu_0, \mu_1 < 1$, $\alpha > 0$ and $\beta \in (0, 1)$ be given constants. Assume that we have a step size $t \geq \min\{\alpha, \beta \bar{t}\}$ such that

$$q_i(s_i(t)) \leq \mu_0 (\!(s_i(t), g_i)\!)_i \quad \text{and} \quad \|s_i(t)\|_i \leq \Delta_i \tag{4.21}$$

holds, where $\bar{t} > 0$ satisfies

$$q_i(s_i(\bar{t})) > \mu_0 (\!(s_i(\bar{t}), g_i)\!)_i \quad \text{or} \quad \|s_i(\bar{t})\|_i \geq \mu_1 \Delta_i.$$

Such a step size exists and can be calculated with a finite number of evaluations of $s(\cdot)$ by a simple backtracking technique similar to the Armijo rule.

Lemma 4.7 *Let Assumption 4.2 and* (4.18) *hold. If $p_i = Proj_{C_i}(-g_i)$ is not smooth, i.e.,* (4.16) *is satisfied, each step $s_i(t)$ where t satisfies* (4.21) *achieves the descent*

$$-q_i(s_i(t)) \geq C^{-1} c(\kappa_\chi, \tau) \chi_i(v_i) \min\left\{\Delta_i, \frac{c(\kappa_\chi, \tau)}{C_H} \chi_i(v_i)\right\}.$$

The constant C is independent of i but depends on the parameters μ_0, μ_1, α and β.

Proof The proof of this lemma runs along the lines of the proof of Theorem 4.4 in [Mor88]. Only simple modifications are necessary; we leave the details to the reader. As in Lemma 4.6 the non-smoothness of the projected gradient must be used to change the norm of p_i. □

4.2.2. Separable constrained problems

From now on for the rest of this chapter, we suppose that the finite dimensional spaces $\mathcal{V}_i$ and feasible sets C_i have a particular structure. We assume that $\mathcal{V}_i$ can be decomposed as

$$\mathcal{V}_i = \sum_{j=1}^{n_i} \mathcal{V}_i^j,$$

such that the decomposition is orthogonal with respect to $(\!(\cdot,\cdot)\!)_i$, i.e.,

$$(\!(v_i^j, v_i^k)\!)_i = 0 \quad \text{for all } v_i^j \in \mathcal{V}_i^j,\ v_i^k \in \mathcal{V}_i^k,\ j \neq k. \tag{4.22}$$

Because of the orthogonality, the representation of an element $v_i \in \mathcal{V}_i$ as $v_i = \sum_{j=1}^{n_i} v_i^j$, $v_i^j \in \mathcal{V}_i^j$, is unique and moreover $|\!|\!| v_i |\!|\!|_i^2 = \sum_{j=1}^{n_i} |\!|\!| v_i^j |\!|\!|_i^2$ holds. In the following, a superscript as in u_i^j denotes the orthogonal projection of u_i onto $\mathcal{V}_i^j$. Notice that n_i need not necessarily be equal to the dimension of $\mathcal{V}_i$ but is allowed to be smaller. Since we will exclusively use it in cases where each $\mathcal{V}_i^j$ is spanned by a single basis vector, we will stick to this notation in the following.

We assume feasible sets $C_i \subset \mathcal{V}_i$ that are the sum of convex subsets of $\mathcal{V}_i^j$, more precisely:

$$C_i = \sum_{j=1}^{n_i} C_i^j, \quad C_i^j \subset \mathcal{V}_i^j \text{ closed and convex.}$$

A simple consequence of the orthogonality of the subspaces $\mathcal{V}_i^j$ and the special structure of C_i is that the projection onto C_i is just the sum of the projections onto each subspace.

Lemma 4.8 *In the setting depicted above, the projection of an element $v_i = \sum_{j=1}^{n_i} v_i^j \in \mathcal{V}_i$ onto the convex set C_i satisfies:*

$$\mathit{Proj}_{C_i}(v_i) = \sum_{j=1}^{n_i} \mathit{Proj}_{C_i^j}(v_i).$$

Proof Let $u_i = \sum_{j=1}^{n_i} u_i^j$, $u_i^j \in C_i^j$. Then

$$\begin{aligned}\Big(\!\Big(v_i - \sum_{k=1}^{n_i} \mathrm{Proj}_{C_i^k}(v_i), \sum_{j=1}^{n_i} \mathrm{Proj}_{C_i^j}(v_i) - u_i\Big)\!\Big)_i &= \sum_{j=1}^{n_i} \Big(\!\Big(v_i - \sum_{k=1}^{n_i} \mathrm{Proj}_{C_i^k}(v_i), \mathrm{Proj}_{C_i^j}(v_i) - u_i^j\Big)\!\Big)_i \\ &= \sum_{j=1}^{n_i} (\!(v_i - \mathrm{Proj}_{C_i^j}(v_i), \mathrm{Proj}_{C_i^j}(v_i) - u_i^j)\!)_i.\end{aligned}$$

From the Projection Theorem A.2, it follows that each summand is non-negative and hence

$$\Big(\!\Big(v_i - \sum_{k=1}^{n_i} \mathrm{Proj}_{C_i^k}(v_i), \sum_{j=1}^{n_i} \mathrm{Proj}_{C_i^j}(v_i) - u_i\Big)\!\Big)_i \geq 0.$$

Now the second part of the Projection Theorem proves the assertion. □

Example 4.1 Consider the setting from Example 3.1. The finite element spaces $\mathcal{V}_i$ can be written as the sum of the one-dimensional spaces that are generated by the nodal basis functions ϕ_i^j,

$$\mathcal{V}_i = \sum_{j=1}^{n_i} \mathcal{V}_i^j, \quad \mathcal{V}_i^j := \{\alpha\phi_i^j \,|\, \alpha \in \mathbb{R}\}.$$

The standard $L^2(\Omega)$-inner product does not satisfy the orthogonality property (3.44). Two common choices (cf. for example [BS08, Sec. 6.2] or [Bra07, Ch. V, Sec. 2]) for an equivalent inner product for $d = 2$ are

$$(\!(u_i, v_i)\!)_i := \frac{1}{3} \sum_{t \in \mathcal{T}_i} |t| (u_i(x_{t,1})v_i(x_{t,1}) + u_i(x_{t,2})v_i(x_{t,2}) + u_i(x_{t,3})v_i(x_{t,3}))$$

or even simpler

$$(\!(u_i, v_i)\!)_i := h_i^d \sum_{x_i^k \in \mathcal{N}_i} u_i(x_i^k)v_i(x_i^k).$$

Here $|t|$ denotes the area of the triangle t and $x_{t,l}$, $l = 1, 2, 3$, its vertices. Since $\phi_i^j(x_i^k) = \delta_{jk}$ for $x_i^k \in \mathcal{N}_i$, it is obvious that both products satisfy the orthogonality assumption (3.44).

In this setting, each closed and convex subset $C_i^j \subset \mathcal{V}_i^j$ can be written as an interval in the coefficient space, i.e., $C_i^j = \{\alpha\phi_i^j \,|\, \alpha \in [\tilde{l}_i^j, \tilde{u}_i^j]\}$ with suitable lower and upper bounds $\tilde{l}_i, \tilde{u}_i \in \mathbb{R}^{n_i}$. Since the elements in $\mathcal{V}_i$ are piecewise linear, the sets

$$C_i = \{v_i \in \mathcal{V}_i \,|\, l_i \le v_i \le u_i\} \quad \text{and} \quad \{v_i \in \mathcal{V}_i \,|\, l_i(x_i^j) \le v_i(x_i^j) \le u_i(x_i^j),\, x_i^j \in \mathcal{N}_i\}$$

coincide for $l_i, u_i \in \mathcal{V}_i$. Hence, we can decompose such sets using the coefficient vectors $\tilde{l}_i$ and $\tilde{u}_i$:

$$C_i = \sum_{j=1}^{n_i} C_i^j, \quad C_i^j := \{\alpha\phi_i^j \,|\, \tilde{l}_i^j \le \alpha \le \tilde{u}_i^j\}.$$

The special form of the inner product makes the projection a cheap operation. Given an element $v_i \in \mathcal{V}_i$ with associated coefficient vector $\tilde{v}_i$, the coefficient vector of the projection $w_i = \mathrm{Proj}_{C_i}(v_i)$ is due to Lemma 4.8 just

$$\tilde{w}_i^j = \min\left\{\tilde{u}_i^j, \max\{\tilde{l}_i^j, \tilde{v}_i^j\}\right\}, \quad j = 1, \ldots, n_i.$$

Notice that the projection of an element in regards to the standard inner product is vastly more expensive.

4.2.3. Smoothers in the strictly convex case

In this section we analyze a *projected* (block) *successive relaxation* algorithm to approximately solve the trust-region subproblem (4.14). Under the additional assumption that the operator H_i is positive definite, we will show that each step satisfies the fraction of Cauchy decrease condition if the smoothness property (2.30) is violated, i.e., (4.16) holds. If $C_i = \mathcal{V}_i$ and $\theta = 1$, the algorithm coincides with the multiplicative subspace correction algorithm, Algorithm 3.2, considered in the previous chapter.

Algorithm 4.1 (PSR)
Choose $\theta \in (0,2)$, set $k=1$ and $y_{i,0}=0$.

Step 1 Find $s_i^{k*} \in C_i^k$ such that

$$q_i(y_{i,k-1}+s_i^{k*}) \le q_i(y_{i,k-1}+u_i^k) \quad \forall\, u_i^k \in C_i^k.$$

Step 2 Set $y_{i,k}=y_{i,k-1}+\theta_k s_i^{k*}$ with $\theta_k=\min\{\theta,\, \max\{t\ge 1 \,|\, y_{i,k-1}+ts_i^{k*}\in C_i\}\}$. If $k<n_i$, set $k \leftarrow k+1$ and go to Step 1.

Step 3 If $\|y_{i,n_i}\|_i > \Delta_i$, set $s_i=\frac{\Delta_i}{\|y_{i,n_i}\|_i}y_{i,n_i}$, otherwise set $s_i=y_{i,n_i}$. Return with s_i.

Notice that the optimization problems in Step 2 possess a unique solution, because q_i is uniformly convex if H_i is positive definite.

Remark 4.3 Algorithm 4.1 can easily be modified to allow more than one optimization sweep through the subspaces:

> Repeat m times: Instead of going to Step 3 when $k=n_i$ holds, restart the algorithm but use $y_{i,0}=y_{i,n_i}$.

Since every step s_i^{k*} produces descent, each y_{i,n_i} has a lower function value than the preceding one.

Remark 4.4 The order in which we process the subspaces in Step 1 can be chosen arbitrarily.

As for unconstrained problems (cf. (3.35)), we impose the sparsity condition

$$\|\gamma_1\|_\infty \le \nu_1 \tag{4.23}$$

with a positive and level-independent constant ν_1. Here, $\gamma_1 \in \mathbb{R}^{n_i\times n_i}$ is the interaction matrix with entries

$$\gamma_1^{jk}=\begin{cases}0 & \text{if } (\!(v_i^j, H_i v_i^k)\!)_i = 0 \text{ for all } v_i^j\in\mathcal{V}_i^j,\ v_i^k\in\mathcal{V}_i^k,\\ 1 & \text{otherwise.}\end{cases}$$

Theorem 4.2 *Let s_i be a step generated by algorithm PSR. Assume that H_i is positive definite and that*

$$(\!(u_i^j, H_i u_i^j)\!)_i \ge (CC_H\lambda_i^{\max})^{-1}|||u_i^j|||_i^2 \quad \textit{for all } u_i^j\in C_i^j,\ j=1,\dots,n_i \tag{4.24}$$

is satisfied. Let furthermore (4.23) *and*

$$\|u_i\|_i^2 \le C\left(\sum_{j=1}^{n_i}(\!(u_i^j, H_i u_i^j)\!)_i\right) \quad \forall u_i\in C_i \tag{4.25}$$

hold. Then s_i is a feasible step of the trust-region subproblem (4.14)*. Moreover, if Assumption 4.2 and* (4.16) *is satisfied, it yields the descent*

$$-q_i(s_i) \geq C^{-1}\frac{2-\theta}{2}c(C_H,\nu_1,\theta)c(\kappa_\chi,\tau)\chi_i(v_i)\min\Big\{\Delta_i,\theta c(\kappa_\chi,\tau)c(C_H,\nu_1,\theta)\chi_i(v_i)\Big\}$$

with $c(C_H,\nu_1,\theta) = \big[\sqrt{C_H}(1+(|1-\theta|+\sqrt{\nu_1}\theta))\big]^{-1}$.

Proof We first show the feasibility. Each partial step s_i^k is element of C_i^k. The definition of θ_k guarantees $\theta_k s_i^k \in C_i^k$ and hence $y_{i,n_i} = \sum_{k=1}^{n_i}\theta_k s_i^{k*} \in C_i$ follows. The scaling in Step 3 ensures $\|s_i\|_i \leq \Delta_i$ and from the convexity of C_i follows $s_i \in C_i$ and thus the feasibility of s_i.

To show the second assertion, we start by estimating the descent achieved by one iteration of Algorithm 4.1. For $1 \leq k \leq n_i$, we have

$$\begin{aligned}
q_i(y_{i,k-1}) - q_i(y_{i,k}) &= -\theta_k(\!(s_i^{k*}, g_i)\!)_i + \frac{1}{2}\Big((\!(y_{i,k}-\theta_k s_i^k, H_i(y_{i,k}-\theta_k s_i^k))\!)_i - (\!(y_{i,k}, H_i y_{i,k})\!)_i\Big)\\
&= \theta_k(\!(s_i^{k*}, g_i + H_i y_{i,k})\!)_i + \frac{\theta_k^2}{2}(\!(s_i^{k*}, H_i s_i^{k*})\!)_i\\
&= -\theta_k(\!(s_i^{k*}, g_i + H_i(y_{i,k-1} + s_i^{k*} - (1-\theta_k)s_i^{k*}))\!)_i + \frac{\theta_k^2}{2}(\!(s_i^k, H_i s_i^k)\!)_i\\
&= -\theta_k(\!(s_i^{k*}, g_i + H_i(y_{i,k-1} + s_i^{k*}))\!)_i + \frac{2\theta_k-\theta_k^2}{2}(\!(s_i^k, H_i s_i^k)\!)_i.
\end{aligned}$$

s_i^{k*} is the solution of the convex optimization problem

$$\min_{u_i^k \in C_i^k}\psi(u_i^k), \quad \psi(u_i^k) := q_i(y_{i,k-1}+u_i^k) \tag{4.26}$$

and hence, due to Lemma 2.2, satisfies the necessary optimality condition (2.24). Since $0 \in C_i^k$, we obtain

$$0 \leq (\!(0 - s_i^{k*}, \nabla\psi(s_i^{k*}))\!)_i = -(\!(s_i^{k*}, g_i + H_i(y_{i,k-1}+s_i^{k*}))\!)_i$$

and thus

$$q_i(y_{i,k-1}) - q_i(y_{i,k}) \geq \frac{2\theta_k-\theta_k^2}{2}(\!(s_i^k, H_i s_i^k)\!)_i.$$

The definition of θ_k implies $\theta_k = \theta$ for $\theta \in (0,1]$ and $1 \leq \theta_k \leq \theta$ for $\theta > 1$. The function $\theta_k \mapsto \frac{1}{2}(2\theta_k - \theta_k^2)$ is monotone decreasing for $\theta_k \in [1,2)$. Hence,

$$q_i(y_{i,k-1}) - q_i(y_{i,k}) \geq \frac{2\theta-\theta^2}{2}(\!(s_i^k, H_i s_i^k)\!)_i$$

is satisfied for all $\theta \in (0,2)$. Representing the difference of the function values as telescope sum yields

$$q_i(0) - q_i(y_{i,n_i}) \geq \frac{2\theta-\theta^2}{2}\sum_{k=1}^{n_i}(\!(s_i^k, H_i s_i^k)\!)_i. \tag{4.27}$$

Taking into account that s_i^{k*} is the optimal solution of (4.26), we get

$$(\!(-\nabla\psi_i(s_i^{k*}), s_i^{k*} - u_i^k)\!)_i \geq 0 \Leftrightarrow (\!((s_i^k - \nabla\psi_i(s_i^{k*})) - s_i^{k*}, s_i^{k*} - u_i^k)\!)_i \geq 0 \quad \forall u_i^k \in C_i^k.$$

The Projection Theorem now shows that

$$s_i^{k*} = \mathrm{Proj}_{C_i^k}(s_i^{k*} - \nabla\psi_i(s_i^{k*})) = \mathrm{Proj}_{C_i^k}(s_i^{k*} - (g_i + H(y_{i,k-1} + s_i^{k*})). \tag{4.28}$$

Using Lemma 4.8, (4.28) and the Cauchy-Schwarz inequality we obtain

$$\begin{aligned}
|||p_i|||_i^2 &= \sum_{k=1}^{n_i} \left(\left(\mathrm{Proj}_{C_i}(-g_i), p_i^k\right)\right)_i \\
&= \sum_{k=1}^{n_i} \left(\left(\mathrm{Proj}_{C_i^k}(-g_i) + \left[s_i^{k*} - \mathrm{Proj}_{C_i^k}(s_i^{k*} - (g_i + H_i(y_{i,k-1} + s_i^{k*}))\right], p_i^k\right)\right)_i \\
&\leq \sum_{k=1}^{n_i} \left[|||s_i^{k*}|||_i + |||\mathrm{Proj}_{C_i^k}(s_i^{k*} - (g_i + H_i(y_{i,k-1} + s_i^{k*}))) - \mathrm{Proj}_{C_i^k}(-g_i)|||_i\right] |||p_i^k|||_i \\
&= \sum_{k=1}^{n_i} \left[|||s_i^{k*}|||_i + |||\mathrm{Proj}_{C_i^k}(s_i^{k*} - (g_i + H_i(y_{i,k-1} + s_i^{k*}))^k) - \mathrm{Proj}_{C_i^k}(-g_i^k)|||_i\right] |||p_i^k|||_i.
\end{aligned}$$

Note the use of $\mathrm{Proj}_{C_i^k}(v_i) = \mathrm{Proj}_{C_i^k}(v_i^k)$ in the last step. From the Lipschitz continuity of the projection (cf. Lemma A.1) follows

$$|||p_i|||_i^2 \leq \sum_{k=1}^{n_i} \left[2|||s_i^{k*}|||_i + |||\left(H_i(y_{i,k-1} + s_i^{k*})\right)^k|||_i\right] |||p_i^k|||_i.$$

Since $y_{i,k}$ is the sum of the (scaled) steps, i.e., $y_{i,k} = \sum_{j=1}^k \theta_j s_i^j$, we further have

$$\begin{aligned}
|||p_i|||_i^2 &\leq 2\left(\sum_{k=1}^{n_i} |||s_i^{k*}|||_i |||p_i^k|||_i\right) + \sum_{k=1}^{n_i} \left|\left|\left|\left[H_i \sum_{j=1}^k \theta_j s_i^j + (1-\theta_j) H_i s_i^{k*}\right]^k\right|\right|\right|_i |||p_i^k|||_i \\
&\leq 2\left(\sum_{k=1}^{n_i} |||s_i^{k*}|||_i |||p_i^k|||_i\right) + \theta \sum_{k=1}^{n_i}\sum_{j=1}^{k} |||(H_i s_i^j)^k|||_i |||p_i^k|||_i + |1-\theta| \sum_{k=1}^{n_i} |||(H_i s_i^{k*})^k|||_i |||p_i^k|||_i.
\end{aligned} \tag{4.29}$$

Note that we have used the triangle inequality and $|1-\theta_j| \leq |1-\theta|$. In the next step, we derive upper bounds for the three sums of the last expression separately.

With the Cauchy-Schwarz inequality, the orthogonality of the decomposition, and (4.24) we can estimate the first term of the sum:

$$2\sum_{k=1}^{n_i} |||s_i^{k*}|||_i |||p_i^k|||_i \leq 2\left(\sum_{k=1}^{n_i} |||s_i^{k*}|||_i^2\right)^{1/2} \left(\sum_{k=1}^{n_i} |||p_i^k|||_i^2\right)^{1/2} \leq \left(CC_H\lambda_i^{\max} \sum_{k=1}^{n_i} (\!(s_i^{k*}, H_i s_i^{k*})\!)_i\right)^{1/2} |||p_i|||_i.$$

Since H_i is positive definite, we conclude from (4.15) that $|||H_i u_i|||_i^2 \leq CC_H\lambda_i^{\max}(\!(u_i, H_i u_i)\!)_i$ for $u_i \in \mathcal{V}_i$. Furthermore, we have

$$|||(H_i s_i^{k*})^k|||^2 \leq \sum_{j=1}^{n_i} |||(H_i s_i^{k*})^j|||^2 = |||H_i s_i^{k*}|||^2.$$

Using the Cauchy-Schwarz inequality, the previous estimate and (4.24), we obtain for the last term:

$$|1-\theta|\sum_{k=1}^{n_i}|||(H_i s_i^{k*})^k|||_i|||p_i^k|||_i \le |1-\theta|\left(\sum_{k=1}^{n_i}|||H_i s_i^{k*}|||_i^2\right)^{1/2}|||p_i|||_i$$
$$\le |1-\theta|\left(CC_H\lambda_i^{\max}\sum_{k=1}^{n_i}(\!(s_i^{k*},H_i s_i^{k*})\!)_i\right)^{1/2}|||p_i|||_i.$$

Assumption (4.23) implies $\sum_{k=1}^{n_i}\gamma_1^{kj}\le\nu_1$ for $1\le j\le n_i$. Thus, similarly to the proof of Lemma 3.10, it follows that

$$\theta\sum_{k=1}^{n_i}\sum_{j=1}^{j}\gamma_1^{kj}|||(H_i s_i^j)^k|||_i|||p_i^k|||_i \le \theta\left(\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}|||(H_i s_i^j)^k|||_i^2\right)^{1/2}\left(\sum_{k=1}^{n_i}\sum_{j=1}^{n_i}\gamma_1^{kj}|||p_i^k|||_i^2\right)^{1/2}$$
$$\le \theta\left(\sum_{k=1}^{n_i}|||H_i s_i^{k*}|||_i^2\right)^{1/2}\left(\sum_{k=1}^{n_i}\nu_1|||p_i^k|||_i^2\right)^{1/2}$$
$$\le \theta\sqrt{\nu_1}\left(CC_H\lambda_i^{\max}\sum_{k=1}^{n_i}(\!(s_i^{k*},H_i s_i^{k*})\!)_i\right)^{1/2}|||p_i|||_i.$$

Combining these estimates with (4.29) gives

$$|||p_i|||_i \le C\left[\sqrt{C_H}(1+(|1-\theta|+\sqrt{\nu_1}\theta))\right]\sqrt{\lambda_i^{\max}}\left(\sum_{k=1}^{n_i}(\!(s_i^k,H_i s_i^k)\!)_i\right)^{1/2}, \tag{4.30}$$

and finally because of (4.27):

$$q_i(0)-q_i(y_{i,n_i}) \ge \frac{2\theta-\theta^2}{2}\frac{1}{C}\left[\sqrt{C_H}(1+(|1-\theta|+\sqrt{\nu_1}\theta))\right]^{-2}\frac{1}{\lambda_i^{\max}}|||p_i|||_i^2. \tag{4.31}$$

If $\|y_{i,n_i}\|_i\le\Delta_i$, the proof is finished after applying

$$\frac{|||p_i|||_i^2}{\lambda_i^{\max}} \ge c(\kappa_\chi,\tau)^2\chi_i(v_i)^2, \tag{4.32}$$

which follows from Assumption 4.2 and (4.16).

Let us now turn to the case $\|y_{i,n_i}\|_i>\Delta_i$. This implies that $s_i=ty_{i,n_i}$ with $t=\frac{\Delta_i}{\|y_{i,n_i}\|_i}<1$. Since H_i is positive definite, we have

$$-q_i(s_i) = -t(\!(g_i,y_{i,n_i})\!)_i-\frac{t^2}{2}(\!(y_{i,n_i},H_i y_{i,n_i})\!)_i \ge -t(\!(g_i,y_{i,n_i})\!)_i-\frac{t}{2}(\!(y_{i,n_i},H_i y_{i,n_i})\!)_i = -tq_i(y_{i,n_i}).$$

From (4.27) and assumption (4.25) follows

$$\begin{aligned}
-q_i(s_i) &\geq \frac{2\theta-\theta^2}{2}\frac{\Delta}{\|y_{i,n_i}\|_i}\sum_{k=1}^{n_i}(\!(s_i^k, H_i s_i^k)\!)_i \\
&\geq \frac{2-\theta}{2}\frac{\Delta}{\|y_{i,n_i}\|_i}\left(\sum_{k=1}^{n_i}(\!(s_i^k, H_i s_i^k)\!)_i\right)^{1/2}\left(\sum_{k=1}^{n_i}(\!(\theta_k s_i^k, H_i\theta_k s_i^k)\!)_i\right)^{1/2} \\
&\geq \frac{2-\theta}{2}\frac{\Delta}{C}\left(\sum_{k=1}^{n_i}(\!(s_i^k, H_i s_i^k)\!)_i\right)^{1/2}.
\end{aligned}$$

Now we use (4.30) and obtain

$$-q_i(s_i) \geq \frac{\Delta}{C}\frac{2-\theta}{2}\left[\sqrt{C_H}(1+(|1-\theta|+\sqrt{\nu_1}\theta))\right]^{-1}(\lambda_i^{\max})^{-1/2}\,|||p_i|||_i.$$

Combining the last estimate with (4.31) and (4.32) we obtain the assertion. □

Remark 4.5 In the special case that the trust-region norm is given by

$$\|u_i\|_i = \max\{\|u_i^k\|_{*,k} \,|\, k = 1,\ldots,n_i\},$$

where $\|\cdot\|_{*,k}$ are arbitrary, we can modify Algorithm PSR to incorporate the trust-region condition in the first step. Instead of seeking the optimal step in the set C_i^k, we consider the set

$$C_i^k \cap \{u_i^k \in \mathcal{V}_i^k \,|\, \|u_i^k\|_{*,k} \leq \Delta_i\}.$$

Obviously, the final iterate y_{i,n_i} then always satisfies $\|y_{i,n_i}\|_i \leq \Delta_i$, and the scaling in Step 3 of the algorithm is not necessary. In the setting of Example 4.1, the $L^\infty(\Omega)$-norm is of this type. We strongly conjecture that a result similar to Theorem 4.2 can be shown for this variant of the algorithm.

Remark 4.6 Instead of this successive algorithm, one can also use a parallel method similar to Algorithm 3.1. The proof of the minimum descent for this variant is similar to the proof of the preceding theorem.

The next lemma shows that (4.24) and (4.25) from the last theorem hold under assumptions that were similarly postulated in Lemma 3.8 and Lemma 3.9 for unconstrained problems.

Lemma 4.9 *1. Let* $\|u_i\|_i \leq C\sqrt{(\!(u_i, H_i u_i)\!)_i}$ *and* (4.23) *be satisfied. Then* (4.25) *holds.*

2. Let $\|u_i\|_i^2 \leq C\lambda_i |||u_i|||_i^2$ *and*

$$(\!(u_i^j, H_i u_i^j)\!)_i \geq \frac{1}{C}\lambda_i |||u_i^j|||_i^2 \quad \text{for all } u_i^j \in C_i^j \text{ and } j = 1,\ldots,n_i \tag{4.33}$$

be satisfied, where λ_i *denotes the largest eigenvalue of* H_i*. Then* (4.25) *and* (4.24) *hold.*

PROOF 1. Since H_i is positive definite and (4.23) is satisfied, we obtain by Lemma 3.10:

$$(\!(u_i, H_i u_i)\!)_i = \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} (\!(u_i^j, H_i u_i^k)\!)_i \leq \nu_1 \sum_{k=1}^{n_i} (\!(u_i^k, H_i u_i^k)\!)_i.$$

(4.25) now follows directly by the assumption on the trust-region norm.

2. Assumption (4.33) implies (4.24). Furthermore,

$$\sum_{k=1}^{n_i} (\!(u_i^k, H_i u_i^k)\!)_i \geq \frac{1}{C} \lambda_i \sum_{k=1}^{n_i} |\!|\!| u_i^k |\!|\!|_i^2 = \frac{1}{C} \lambda_i |\!|\!| u_i |\!|\!|_i^2 \geq \frac{1}{C} \|u_i\|_i^2$$

holds, which shows (4.25). □

4.2.4. Non-convex trust-region subproblems

If H_i is not positive definite, we cannot prove the fraction of Cauchy decrease condition for a step generated by the PSR algorithm (cf. Example 3.4). Since the simple projected gradient step does not have a good smoothing effect, we combine both algorithms by a strategy motivated by the classical Dogleg method due to [Pow70] for the approximate solution of trust-region subproblems.

Algorithm 4.2 (DoglegSmoothing(m))
Choose $\theta \in (0, 2)$, set $j = 1$ and $k = 1$.

Step 1 Calculate a solution t^* of problem (4.17) and set $\hat{s}_i = s_{i,C} = t^* \mathrm{Proj}_{C_i}(-g_i)$.

Step 2 Find $s_i^{k*} \in C_i^k \cap B_i^k$, where $B_i^k \subset \mathcal{V}_i^k$ is an arbitrary compact set, such that

$$q_i(\hat{s}_i + s_i^{k*}) \leq q_i(\hat{s}_i + u_i^k) \quad \forall\, u_i^k \in C_i^k \cap B_i^k.$$

Step 3 Update $\hat{s}_i \leftarrow \hat{s}_i + \theta_k s_i^{k*}$ with $\theta_k = \min\{\theta, \max\{t \geq 1 \,|\, \hat{s}_i + t s_i^{k*} \in C_i\}\}$. If $k = n_i$ and $j = m$ go to Step 4, otherwise set $k \leftarrow k + 1$ if $k < n_i$ or $k = 1$ and $j \leftarrow j + 1$ if $k = n_i$. Go to Step 2.

Step 4 Find the solution t_i^* of the trust-region subproblem (4.14) on the path

$$s(t) = s_{i,C} + t(\hat{s}_i - s_{i,C}), \quad t \in [0, 1],$$

and return with $s_i^* = s(t^*)$.

The compact sets B_i^k are needed to ensure the solvability of the problems in Step 2, since the sets C_i^k need not be compact. One possibility is to use the trust-region condition by setting $B_i^k = \{u_i^k \in \mathcal{V}_i^k \,|\, \|u_i^k\|_i \leq \Delta_i\}$. From Lemma 4.6 it follows that the step $s_{i,C}$ generated by Algorithm 4.2 satisfies the fraction of Cauchy decrease condition under the assumption postulated in the lemma. Since the descent produced by the final step s_i^* is even larger, it also satisfies (2.29).

Remark 4.7 In our numerical implementation we use a slightly different version of the above algorithm: If $\|\hat{s}_i\|_i > \Delta_i$ in Step 4, we also calculate the point $\bar{s}_i = \Delta_i / \|\hat{s}_i\|_i \hat{s}_i$, and return with $\bar{s}_i$ if $q_i(\bar{s}_i) < q_i(s_i^*)$ holds.

It should be possible to construct a pure subspace correction algorithm, similar to Algorithm 3.3, if H_i is not positive definite. However, even in the unconstrained case, our numerical tests show that Algorithm 4.2 is superior to Algorithm 3.3, and hence we have not studied it further for unconstrained problems.

4.3. Construction of lower-level boxes

Both, for the trust-region algorithm (Algorithm 2.1) and the multilevel stationarity measure χ_i^{ML}, it is necessary to construct feasible sets on the lower-levels. Let k be the current iteration index of Algorithm 2.1 on level i and $j \in N(i)$. Before we enter level j in step 2, we have to construct a closed and convex set $C_j(v_{i,k})$ that satisfies

$$0 \in C_j(v_{i,k}) \quad \text{and} \quad v_{i,k} + P_j^i s_j \in C_i \text{ for all } s_j \in C_j(v_{i,k}). \tag{4.34}$$

It is favorable that the lower-level set $C_j(v_{i,k})$ has the same structure as the feasible set C_i. In order to simplify notation, we assume without loss of generality $v_{i,k} = 0$.

In this section we consider a special class of convex sets that occurs frequently, and show how to construct suitable lower-level sets.

Let $\{\phi_i^k\}_{k=1}^{n_i} \subset \mathcal{V}_i$ be a basis of $\mathcal{V}_i$. We call a set C_i a *box* if there exist $\tilde{l}_i \in \{\mathbb{R} \cup \{-\infty\}\}^{n_i}$ and $\tilde{u}_i \in \{\mathbb{R} \cup \{\infty\}\}^{n_i}$ such that

$$C_i = \left\{ v_i \in \mathcal{V}_i \,|\, v_i = \sum_{k=1}^{n_i} \tilde{v}_i^k \phi_i^k, \quad \tilde{l}_i \le \tilde{v}_i \le \tilde{u}_i \right\}. \tag{4.35}$$

Here and in the following, the inequality between two vectors is applied component wise. Note that these are just the type of sets we have considered in Example 4.1.

We will present two possibilities for the construction of lower-level boxes. We make the analysis using an arbitrary transfer operator $T_j^i \colon \mathcal{V}_j \to \mathcal{V}_i$ instead of the prolongation P_j^i. This is because we will introduce a slight variation of the algorithm in Section 4.3.2, where P_j^i is replaced by a modified prolongation operator.

We denote by $\tilde{T}_j^i \in \mathbb{R}^{n_i \times n_j}$ the matrix representation of T_j^i that operates on the coefficient vectors. That means, given the bases $\{\phi_i^k\}$ and $\{\phi_j^k\}$ of $\mathcal{V}_i$ and $\mathcal{V}_j$, we have the identity

$$T_j^i v_j = \sum_{k=1}^{n_i} (\tilde{T}_j^i \tilde{v}_j)^k \phi_i^k, \quad \text{for all } v_j = \sum_{k=1}^{n_j} \tilde{v}_j^k \phi_j^k.$$

Throughout this section we always use a tilde to denote the associated coefficient vector.

Let C_i be a box with bound vectors $\tilde{l}_i \in \{\mathbb{R} \cup \{-\infty\}\}^{n_i}$ and $\tilde{u}_i \in \{\mathbb{R} \cup \{\infty\}\}^{n_i}$. Together with the prior assumptions, condition (4.34) translates as follows: Seek $\tilde{l}_j \in \{\mathbb{R} \cup \{-\infty\}\}^{n_j}$ and $\tilde{u}_j \in \{\mathbb{R} \cup \{\infty\}\}^{n_j}$ such that

$$\tilde{l}_j \leq 0 \leq \tilde{u}_j \quad \text{and} \quad \tilde{l}_i \leq \tilde{T}_j^i \tilde{v}_j \leq \tilde{u}_i \text{ for all } \tilde{v}_j \in [\tilde{l}_j, \tilde{u}_j]. \tag{4.36}$$

The first lemma describes a well-known construction that can be found in various papers, e.g., [GMTWM08, Man84, Kor97, Tai03]. For notational clarity, we often omit the level indices of the transfer operators in the remainder of the section. By $\tilde{T}^{lk}$ we denote the entry in row l and column k of $\tilde{T}$.

Lemma 4.10 *Let C_i a box of the form (4.35) that satisfies $0 \in C_i$. Let furthermore $j \in N(i)$ and $T \colon \mathcal{V}_j \to \mathcal{V}_i$ be a linear transfer operator with corresponding matrix $\tilde{T}$. Assume $\tilde{l}_j$ and $\tilde{u}_j$ are defined by*

$$\tilde{u}_j^k := \min \left\{ \infty, \min_{\substack{m=1,\dots,n_i \\ \tilde{T}^{mk}>0}} \left\{ \frac{\tilde{u}_i^m}{\tau_m} \right\}, \min_{\substack{m=1,\dots,n_i \\ \tilde{T}^{mk}<0}} \left\{ \frac{-\tilde{l}_i^m}{\tau_m} \right\} \right\}, \tag{4.37a}$$

$$\tilde{l}_j^k := \max \left\{ -\infty, \max_{\substack{m=1,\dots,n_i \\ \tilde{T}^{mk}>0}} \left\{ \frac{\tilde{l}_i^m}{\tau_m} \right\}, \max_{\substack{m=1,\dots,n_i \\ \tilde{T}^{mk}<0}} \left\{ \frac{-\tilde{u}_i^m}{\tau_m} \right\} \right\} \tag{4.37b}$$

for $k = 1, \dots, n_j$, where $\tau_m := \sum_{k=1}^{n_j} |\tilde{T}^{mk}|$. Then $\tilde{l}_j$ and $\tilde{u}_j$ satisfy (4.36).

Proof Let $m \in \{1, \dots, n_i\}$ arbitrary. A simple calculation using (4.37a) and (4.37b) shows that $(\tilde{T}\tilde{s}_j)^m = \tilde{T}^l \tilde{s}_j \leq \tilde{u}_i^m$ holds:

$$\begin{aligned}(\tilde{T}\tilde{s}_j)^m &= \sum_{k=1}^{n_j} \tilde{T}^{mk} \tilde{s}_j^k = \sum_{\substack{k=1,\dots,n_j \\ \tilde{T}^{mk}>0}} \tilde{T}^{mk} \tilde{s}_j^k + \sum_{\substack{k=1,\dots,n_j \\ \tilde{T}^{mk}<0}} \tilde{T}^{mk} \tilde{s}_j^k \\ &\leq \sum_{\substack{k=1,\dots,n_j \\ \tilde{T}^{mk}>0}} \tilde{T}^{mk} \frac{\tilde{u}_i^m}{\tau_m} + \sum_{\substack{k=1,\dots,n_j \\ \tilde{T}^{mk}<0}} (-\tilde{T}^{mk}) \frac{\tilde{u}_i^m}{\tau_m} = \frac{\tilde{u}_i^m}{\tau_m} \sum_{k=1}^{n_j} |\tilde{T}^{mk}| = \tilde{u}_i^m\end{aligned}$$

In the same way one shows that $(\tilde{T}\tilde{s}_j)^l \geq \tilde{l}_i^l$. This establishes the second condition in (4.36). The first condition is a direct consequence of the assumption $0 \in C_i$. □

Although the result seems rather technical at first glance, the next example shows in a concrete case that the construction of the bounds is quite natural.

Example 4.2 We assume the setting of Example 3.1, where the spaces $\mathcal{V}_i$ consists of continuous and piecewise linear functions. Since we have a nodal basis $\{\phi_i^k\}_{k=1,\dots,n_i}$ with $\phi_i^k(x_i^l) = \delta_{kl}$ for all $x_i^l \in \mathcal{N}_i$, a box C_i can be written as

$$C_i = \left\{ v_i \in \mathcal{V}_i \,|\, \tilde{l}_i^l \leq v_i(x_i^l) \leq \tilde{u}_i^l \ \forall l = 1, \dots, n_i \right\}.$$

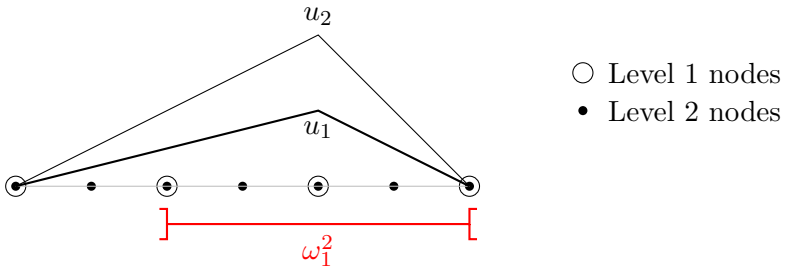

Figure 4.1.: Example of a smooth upper bound u_2 and its lower-level approximation according to (4.38).

We assume that $\tilde{l}_i^l > -\infty$ and $\tilde{u}_i^l < \infty$ for all $l = 1, \ldots, n_i$. Then C_i satisfies

$$C_i = \{v_i \in \mathcal{V}_i \,|\, l_i \leq v_i \leq u_i\}$$

with $l_i, u_i \in \mathcal{V}_i$, because the finite elements are piecewise linear.

Now let $j \in N(i)$ and T_j^i be the identity from $\mathcal{V}_j \to \mathcal{V}_i$. The support of a given basis function ϕ_j^k is given by

$$\omega_j^k = \{x \in \Omega \,|\, \phi_j^k(x) \neq 0\}.$$

It is easy to see that the entries of $\tilde{l}_j$ and $\tilde{u}_j$ given by (4.37a) and (4.37b) satisfy

$$\begin{aligned} \tilde{l}_j^k &= \max\left\{l_i(x_i) \,|\, x_i \in \mathcal{N}_i \cap \omega_j^k\right\}, \\ \tilde{u}_j^k &= \min\left\{u_j(x_i) \,|\, x_i \in \mathcal{N}_i \cap \omega_j^k\right\}. \end{aligned} \tag{4.38}$$

It is clear that these bounds satisfy the necessary conditions.

In the case of pointwise bounds, we can interpret the procedure of constructing u_j as application of a nonlinear operator $I_j^{\ominus}: \mathcal{V}_r \to \mathcal{V}_j$ to the upper bound u_i. Similarly, we can denote the construction of the lower bound by an operator $I_j^{\oplus}$. These operators are exactly the ones analyzed by Tai in [Tai03].

This approach of obtaining lower-level bounds is cheap to calculate but has a serious disadvantage: If the bounds on the fine grid are smooth in the sense that $u_i \in \mathcal{V}_j$ resp. $l_i \in \mathcal{V}_j$, the bounds created by (4.38) are in general too pessimistic (cf. Figure 4.1 for an example).

We next describe a more advanced construction of the boxes, which uses a successive approach to determine the lower-level box. The method is a generalization of a construction proposed by Kornhuber in [Kor97, Sec. 3.1.3], who uses it in the case of linear finite elements.

For the sake of clarity, we restrict ourselves to the case of transfer matrices $\tilde{T}_j^i$ whose entries are non-negative, which is the case when using nested linear finite element spaces or finite differences and the typical bilinear interpolation as prolongation. The extension to more general transfer operators is possible but more technical.

Algorithm 4.3 (CreateCorseBounds($\tilde{u}_i$, $\tilde{T}$))
Step 0 Create an initial guess $u_j \in \mathcal{V}_j$ with $\tilde{u}_j \geq 0$. Set $m = 1$.

Step 1 If $\sum_{k=1}^{n_j} \tilde{T}^{mk}\tilde{u}_j^k \leq \tilde{u}_i^m$, go to Step 4.

Step 2 Set $\tau_m := \sum_{k=1}^{n_j} \tilde{T}^{mk}$. Define the index sets

$$I_{\leq} := \{k = 1, \ldots, n_j \,|\, \tilde{T}^{mk} > 0,\ \tau_m\tilde{u}_j^k \leq \tilde{u}_i^m\},$$
$$I_{>} := \{k = 1, \ldots, n_j \,|\, \tilde{T}^{mk} > 0,\ \tau_m\tilde{u}_j^k > \tilde{u}_i^m\}.$$

Step 3 For $k \in I_{>}$ set

$$\tilde{u}_j^k = \min\left\{\tilde{u}_j^k, \Big(\tilde{u}_i^m - \sum_{l\in I_{\leq}} \tilde{T}^{ml}\tilde{u}_j^l\Big)\Big(\sum_{l\in I_{>}} \tilde{T}^{ml}\Big)^{-1}\right\}. \tag{4.39}$$

Step 4 If $m < n_i$, set $m \leftarrow m + 1$ and go to Step 2. Otherwise return with $\tilde{u}_j$.

Remark 4.8 The algorithm can also be used to calculate a lower bound vector $\tilde{l}_j$ by starting it with $-\tilde{l}_i$ instead of $\tilde{u}_i$ and by setting $\tilde{l}_j = -\text{CreateCorseBounds}(-\tilde{l}_i, j, \tilde{T})$.

Remark 4.9 The first guess in Step 0 can be chosen arbitrarily. One obvious choice is to use the restriction operator R_i^j if $\mathcal{W}_j = \mathcal{V}_j$ holds.

In Figure 4.2 the execution of Algorithm 4.3 is demonstrated on a simple one dimensional example. The identity is used as transfer operator. The corresponding matrix satisfies $\tau_k = 1$ for all $k = 1, \ldots, n_i$. Given the upper bound u_2, the algorithm starts with a lower-level approximation u_1 of the bound u_2. In the example we use the pointwise interpolant. For $k \in \{1, 3, 5, 7\}$ the algorithm does not enter Step 2 and 3 since the values at the nodes already satisfy $u_1(k) \leq u_2(k)$. For $k = 2$ we have $I_{\leq} = \{1\}$ and $I_{>} = \{3\}$. The value of u_1 at $x = 3$ is reduced such that $u_1(2) = u_2(2)$. The same happens for $k = 4$ with $u_1(5)$. For $k = 6$, we have $u_1(6) < u_2(6)$ and hence we do not enter Step 2. The function $\hat{u}_1$ in the last graph is the upper bound obtained by the construction in Lemma 4.10, which is far more pessimistic than the bound u_1 obtained by Algorithm 4.3.

The next lemma proves that Algorithm 4.3 creates suitable lower-level bounds.

Lemma 4.11 *Let C_i be a box of the form* (4.35) *that satisfies $0 \in C_i$. Let $j \in N(i)$ and $T\colon \mathcal{V}_j \to \mathcal{V}_i$ be a linear transfer operator with associated matrix $\tilde{T}$ whose entries are non-negative. Let the coefficient vectors $\tilde{u}_j$ and $\tilde{l}_j$ be generated by Algorithm 4.3 (see also Remark 4.8). Then $\tilde{l}_j$ and $\tilde{u}_j$ satisfy* (4.36).

Proof Let $\tilde{u}_j$ be the coefficient vector in Step 4 of iteration m. Since (4.39) holds for all $k \in I_{>}$,

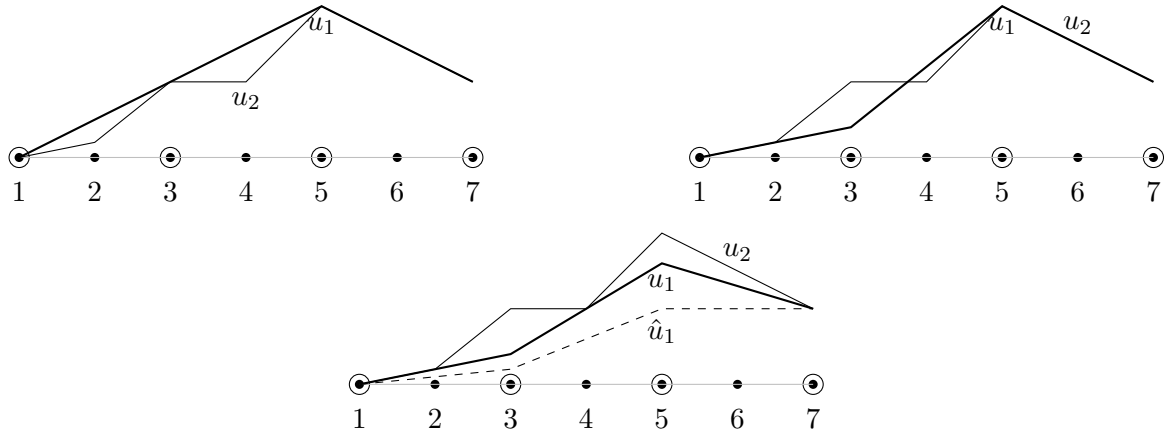

Figure 4.2.: Successive approach of determining the lower-level bound

a simple calculation yields

$$\begin{aligned}(\tilde{T}\tilde{u}_j)^m &= \sum_{k\in I_\leq} \tilde{T}^{mk}\tilde{u}_j^k + \sum_{k\in I_>} \tilde{T}^{mk}\tilde{u}_j^k \leq \sum_{k\in I_\leq} \tilde{T}^{mk}\tilde{u}_j^k + \sum_{k\in I_>} \left[\tilde{T}^{mk}\Big(\tilde{u}_i^m - \sum_{l\in I_\leq} \tilde{T}^{ml}\tilde{u}_j^l\Big)\Big(\sum_{l\in I_>} \tilde{T}^{ml}\Big)^{-1}\right] \\ &\leq \sum_{k\in I_\leq} \tilde{T}^{mk}\tilde{u}_j^k + \Big(\tilde{u}_i^m - \sum_{l\in I_\leq} \tilde{T}^{ml}\tilde{u}_j^l\Big)\Big(\sum_{l\in I_>} \tilde{T}^{ml}\Big)^{-1}\Big(\sum_{k\in I_>} \tilde{T}^{mk}\Big) = \tilde{u}_i^m.\end{aligned}$$

Obviously, the algorithm never increases the entries of $\tilde{u}_j$. So, if $(\tilde{T}\tilde{u}_j)^m \leq u_i^m$ holds after iteration m, it is also satisfied for the final vector, which follows from the non-negativity of $\tilde{T}$. This shows $\tilde{T}\tilde{u}_j \leq \tilde{u}_i$. The lower-level bound $\tilde{l}_j$ is the result of $-\text{CreateCorseBounds}(-\tilde{l}_i, j, \tilde{T})$ and hence $(-\tilde{T}\tilde{l}_j)^m \leq -\tilde{l}_i^m \Leftrightarrow (\tilde{T}\tilde{l}_j)^m \geq \tilde{l}_i^m$ holds. This proves the second condition in (4.36).

It is left to show that $\tilde{l}_j \leq 0 \leq \tilde{u}_j$. The first guess of $\tilde{u}_j$ in Step 0 is non negative and by assumption $\tilde{u}_i \geq 0$ holds. From the definition of $I_\leq$ and τ_m follows

$$\sum_{k\in I_\leq} \tilde{T}^{mk}\tilde{u}_j^k \leq \frac{\tilde{u}_i^m}{\tau_m} \sum_{k\in I_\leq} \tilde{T}^{mk} \leq \tilde{u}_i^m.$$

Hence,

$$\Big(\tilde{u}_i^m - \sum_{k\in I_\leq} \tilde{T}^{mk}\tilde{u}_j^k\Big)\Big(\sum_{k\in I_>} \tilde{T}^{mk}\Big)^{-1} \geq 0$$

and thus

$$\min\left\{\tilde{u}_j^m, \Big(\tilde{u}_i^m - \sum_{k\in I_\leq} \tilde{T}^{mk}\tilde{u}_j^k\Big)\Big(\sum_{k\in I_>} \tilde{T}^{mk}\Big)^{-1}\right\} \geq 0.$$

This shows that after each iteration all entries in $\tilde{u}_j$ are non-negative. As a consequence of the construction of $\tilde{l}_j$ this also proves $-\tilde{l}_j \geq 0$, which completes the proof. □

Remark 4.10 Note that the proof of the algorithm does not depend on the order in which the entries k are processed. The final bound $\tilde{u}_j$, however, is in general different when choosing a different order.

Finally, we show that the bounds obtained by Algorithm 4.3 are at least as good as the bounds defined by (4.37a)–(4.37b).

Lemma 4.12 *Let the assumptions of Lemma 4.11 hold. If the initial guess in Step 0 of Algorithm 4.3 satisfies*

$$\tilde{u}_j^k \geq \min_{\substack{m=1,\dots,n_i \\ \tilde{T}^{mk}>0}} \left\{ \frac{\tilde{u}_i^m}{\tau_m} \right\} \quad \text{for all } k = 1, \dots, n_j, \tag{4.40}$$

then it also holds for the final bound.

Proof We prove this result by induction. From the assumption it follows that the inequality (4.40) holds for $\tilde{u}_j$ in Step 1 in the first iteration of Algorithm 4.3. Now, assume that the assertion is true for iteration m. The only place where $\tilde{u}_j$ is altered is in Step 3. From the definition of the set $I_{\leq}$ we obtain

$$\tilde{u}_i^m - \sum_{k \in I_{\leq}} \tilde{T}^{mk} \tilde{u}_j^k \geq \tilde{u}_i^m - \frac{\tilde{u}_i^m}{\tau_m} \sum_{k \in I_{\leq}} \tilde{T}^{mk} = \frac{\tilde{u}_i^m}{\tau_m} \sum_{k \in I_{>}} \tilde{T}^{mk}.$$

Hence, it follows from the induction hypothesis that the new $\tilde{u}_j^k$, $k \in I_{>}$, after Step 3 still satisfy

$$\tilde{u}_j^k \geq \frac{\tilde{u}_i^m}{\tau_m} \geq \min_{\substack{m=1,\dots,n_i \\ \tilde{T}^{mk}>0}} \left\{ \frac{\tilde{u}_i^m}{\tau_m} \right\}.$$

Therefore, the assertion is true for $m \leftarrow m+1$. This finishes the proof. □

4.3.1. Uniform continuity of χ_i^{ML}

We now come back to the question how strong the continuity of the multilevel stationarity measure χ_i^{ML} depends on the level. We consider the typical setting of Example 4.2 where the feasible set is given by a box $C_i := \{v_i \in \mathcal{V}_i \,|\, l_i \leq v_i \leq u_i\}$ with lower and upper bounds $l_i, u_i \in \mathcal{V}_i$.

As already mentioned in Example 4.2, Tai studied in [Tai03] the interpolation operators $I_j^{\ominus}$ and $I_j^{\oplus}$. We recall that the lower-level feasible sets $C_j(v_i)$ using the bounds of Lemma 4.10 can be written as

$$C_j(v_i) = \{v_j \in \mathcal{V}_j \,|\, l_j := I_j^{\oplus}(l_i - v_i) \leq v_j \leq I_j^{\ominus}(u_i - v_i)\}. \tag{4.41}$$

The following error estimate are shown in [Tai03, Theorem 2]:

Theorem 4.3 *For any $v_i, w_i \in \mathcal{V}_i \subset H^1(\Omega)$ it holds*

$$\begin{aligned} \|I_j^{\ominus}(v_i) - I_j^{\ominus}(w_i) - (v_i - w_i)\|_{L^2(\Omega)} &\leq c_d h_j |v_i - w_i|_{H^1(\Omega)}, \\ \|I_j^{\oplus}(v_i) - I_j^{\oplus}(w_i) - (v_i - w_i)\|_{L^2(\Omega)} &\leq c_d h_j |v_i - w_i|_{H^1(\Omega)} \end{aligned}$$

with $c_d = C$ if $d = 1$, $c_d = C(1 + |\log h_j/h_i|^{1/2})$ if $d = 2$ and $c_d = C(h_j/h_i)^{1/2}$ if $d = 3$.

The next corollary follows directly by using the inverse triangle inequality:

Corollary 4.2 *For any $v_i, w_i \in \mathcal{V}_i \subset H^1(\Omega)$ it holds*

$$\|I_j^{\ominus}(v_i) - I_j^{\ominus}(w_i)\|_{L^2(\Omega)} \leq (C + c_d h_j)\|v_i - w_i\|_{H^1(\Omega)},$$
$$\|I_j^{\oplus}(v_i) - I_j^{\oplus}(w_i)\|_{L^2(\Omega)} \leq (C + c_d h_j)\|v_i - w_i\|_{H^1(\Omega)}$$

with c_d as in Theorem 4.3.

The Hausdorff distance of two boxes can easily be written in terms of the bounds:

Lemma 4.13 *Let $S_k := \{v \in \mathcal{V}_i \,|\, l_k \leq v \leq u_k\}$, $k = 1, 2$, be nonempty sets with $l_k \leq u_k$. Then the following estimate holds:*

$$d_H(S_1, S_2) \leq \left(|||l_1 - l_2|||_i^2 + |||u_1 - u_2|||_i^2\right)^{1/2},$$

where the Hausdorff distance is measured with respect to $|||\cdot|||_i$.

Proof Let $v_1 \in S_1$ be arbitrary. The distance to the set S_2 is given by

$$d(v_1, S_2) = |||\mathrm{Proj}_{S_2}(v_1) - v_1|||_i.$$

In the following we set $v_2^* := \mathrm{Proj}_{S_2}(v_1)$. Using Lemma 4.8 and the orthogonality of the subspaces $\mathcal{V}_i^j := \{\alpha\phi_i^j | \alpha \in \mathbb{R}\}$ with respect to $(\!(\cdot,\cdot)\!)_i$ it follows:

$$d(v_1, S_2)^2 = \sum_{j=1}^{n_i} |||\mathrm{Proj}_{S_1^j}(v_1) - v_1^j|||_i^2 = \sum_{j=1}^{n_i} |(\tilde{v}_2^*)^j - \tilde{v}_1^j|^2 |||\phi_i^j|||_i^2.$$

If $\tilde{l}_2^j \leq \tilde{v}_1^j \leq \tilde{u}_2^j$, we obviously have $0 = |(\tilde{v}_2^*)^j - \tilde{v}_1^j| \leq \max\{|\tilde{l}_2^j - \tilde{l}_1^j|, |\tilde{u}_2^j - \tilde{u}_1^j|\}$. If $(\tilde{v}_2^*)^j < \tilde{v}_1^j$, then $(\tilde{v}_2^*)^j = \tilde{u}_2^j$ holds and because of $\tilde{v}_1^j \leq \tilde{u}_1^j$ we can estimate $|(\tilde{v}_2^*)^j - \tilde{v}_1^j| \leq |\tilde{u}_2^j - \tilde{u}_1^j|$. Similarly, if $(\tilde{v}_2^*)^j > \tilde{v}_1^j$ we have $|(\tilde{v}_2^*)^j - \tilde{v}_1^j| \leq |\tilde{l}_2^j - \tilde{l}_1^j|$. Thus, we obtain for the distance the estimate

$$d(v_1, S_2)^2 \leq \sum_{j=1}^{n_i} \max\{|\tilde{l}_2^j - \tilde{l}_1^j|, |\tilde{u}_2^j - \tilde{u}_1^j|\}^2 |||\phi_i^j|||_i^2 \leq |||l_2 - l_1|||_i^2 + |||u_2 - u_1|||_i^2.$$

By the same argumentation, we obtain the identical bound for $d(S_1, v_2)^2$, $v_2 \in S_2$. This finishes the proof. □

Using the previous lemma, we can estimate the Hausdorff distance of two sets $C_j(v_i)$ and $C_j(w_i)$ which were generated according to Lemma 4.10 by

$$d_H(C_j(v_i), C_j(w_i))^2 \leq |||I_j^{\ominus}(u_i - v_i) - I_j^{\ominus}(u_i - w_i)|||_j^2 + |||I_j^{\oplus}(l_i - v_i) - I_j^{\oplus}(l_i - w_i)|||_j^2.$$

From the level-independent equivalence of the norms $\|\cdot\|$ and $|||\cdot|||_i$ on $\mathcal{V}_j$ and Corollary 4.2, it further follows

$$d_H(C_j(v_i), C_j(w_i))^2 \leq (C + c_d h_j)^2 \|v_i - w_i\|_{\mathcal{V}_i}^2.$$

We recall that this estimates shows that assumption (4.9) of Theorem 4.1 is satisfied with $c_j = (1 + c_d h_j)$.

To analyze the amount of which the constants $\delta(\varepsilon)$ in Corollary 4.1 depends on the level i and the meshsize h_i, we assume that h_i' is uniformly continuous and that the mapping $\varepsilon_g \mapsto \delta(\varepsilon_g)$ is level-independent. Furthermore, we assume that also the constant β_i does not depend on i. These assumptions are not very restrictive if we suppose that the functions f_i are discrete versions of a uniformly continuous differentiable non-linear functional $f\colon \mathcal{V} \to \mathbb{R}$, whose derivatives are bounded. If this is not satisfied, we would also not expect that the f_i and the derived lower-level models h_i are having this features (level-independently). Note that for the global convergence proof of the trust-region algorithm (Theorem 2.2) we only need the uniform continuity of χ_r, which is independent of the concrete choice of the lower-level models.

Under these assumptions, the only term left that depends on the level is the constant B_i which was defined in Theorem 4.1 by

$$B_i^2 := \max\Big\{1, \sum_{j=1}^{i} \big((\lambda_j^{\max})^{-1} c_j^2\big)\Big\}.$$

We recall that in the current setting, $\lambda_j^{\max} \geq C^{-1} h_j^{-2}$ holds and hence $B_i^2 - 1 \leq C \sum_{j=1}^{i} h_j^2 c_j^2$ is satisfied.

We assume that there exists a constant $\gamma < 1$ with $h_j \approx \gamma h_{j-1}$ for all $j = 2, \ldots, r$. We typically have $\gamma = 1/2$ in the case of uniform refinement. Thus we get $c_j = (1 + c_d h_1 \gamma^{j-1})$ and

$$\begin{aligned} B_i^2 - 1 &\leq C \sum_{j=1}^{i} h_j^2 c_j^2 \leq C h_1^2 \Big(\sum_{j=1}^{i} h_1^2 c_d^2 \gamma^{4(j-1)} + \sum_{j=1}^{i} \gamma^{2(j-1)} \Big) \\ &\leq C h_1^4 \sum_{j=1}^{i} c_d^2 \gamma^{4(j-1)} + C h_1^2 (1 - \gamma^2)^{-1}. \end{aligned}$$

The second term does not depend on i, therefore we only consider the first term in the following. Since c_d is constant for $d = 1$, the sum in the first term is bounded by means of the geometric series. Hence, B_i^2 and thus also the uniform continuity does not depend on i or on the meshsize h_i. For $d > 1$ we obtain a weak dependence on i:

d=2:

$$\begin{aligned} C h_1^4 \sum_{j=1}^{i} c_d^2 \gamma^{4(j-1)} &\leq C h_1^4 \sum_{j=1}^{i} (1 + \log |h_j / h_i|) \gamma^{4(j-1)} \\ &\leq C h_1^4 \sum_{j=1}^{i} (1 + i - j) \gamma^{4(j-1)} \leq C h_1^4 i (1 - \gamma^4)^{-1} \end{aligned}$$

d=3:

$$\begin{aligned} C h_1^4 \sum_{j=1}^{i} c_d^2 \gamma^{4(j-1)} &\leq C h_1^4 \sum_{j=1}^{i} h_j / h_i \gamma^{4(j-1)} = C h_1^5 (h_i)^{-1} \sum_{j=1}^{i} \gamma^{5(j-1)} \\ &\leq C h_1^5 h_i^{-1} (1 - \gamma^5)^{-1} \end{aligned}$$

Although we have a dependency on the number of levels used, it is rather weak. If we consider the estimates derived in Theorem 4.1 respectively Corollary 4.1 we have the bounds

$$\delta(\varepsilon) \leq C\varepsilon^4\beta_i^3 h_1^{-4} i^{-1} \text{ if } d = 2 \quad \text{and} \quad \delta(\varepsilon) \leq C\varepsilon^4\beta_i^3 h_1^{-5}\gamma^i \text{ if } d = 3.$$

Remark 4.11 Unfortunately, we cannot show a result like Corollary 4.2 for the construction according to Algorithm 4.3. This comes from the fact that a small difference in one point can propagate and leads to completely different lower level bounds. As an example consider the functions in Figure 4.11. Although the bounds u_2 and $\bar{u}_2$ differ only in the node x_2, the resulting lower-level bounds u_1 and $\bar{u}_1$ are different in any coarse grid point. A straightforward calculation, using an example like this, shows that the results of Corollary 4.2 do not hold.

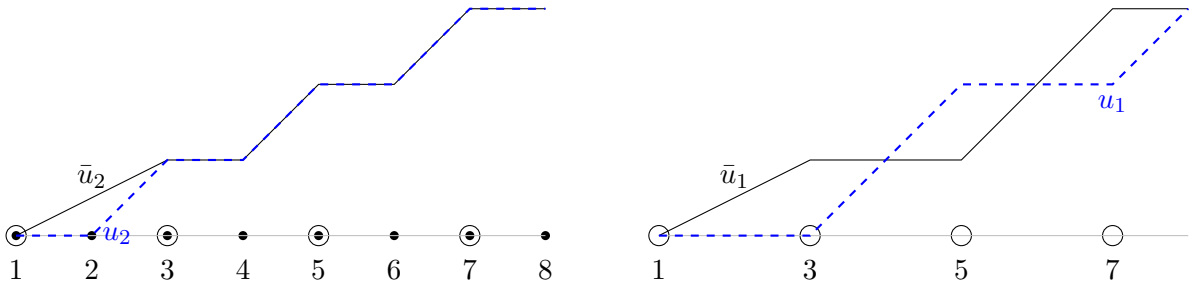

Figure 4.3.: Construction of two lower-level bounds by Algorithm 4.3

However, we can use these lower-level bounds for the calculation of the step in the lower-level trust-region subproblem.

4.3.2. Active sets

Even the best algorithms for constructing lower-level bounds will not succeed in providing a good approximation if the bounds are oscillatory (cf. Figure 2.5). Another source of poor lower-level bounds are active fine grid components. We call an index $m \in \{1, \ldots, n_i\}$ of a coefficient vector $\tilde{v}_i \in \mathbb{R}^{n_i}$ *active* if $\tilde{v}_i^l = \tilde{u}_i^l$ or $\tilde{v}_i^l = \tilde{l}_i^l$ holds. If $\mathcal{V}_i$ is equipped with a nodal basis, each active coefficient $\tilde{v}_i^l$ corresponds to an active node $x_i^l \in \mathcal{N}_i$, i.e., $v_i(x_i^l) = l_i(x_i^l)$ or $v_i(x_i^l) = u_i(x_i^l)$ is satisfied.

Why active components can lead to small feasible sets on lower levels is illustrated on a simple multilevel example in one dimension using piecewise linear functions (Figure 4.4). The upper bound u_3 on level 3 is active at $x = 2$ in this example. Due to this, every feasible step s_i, $i = 1, 2, 3$, must satisfy $s_i(2) \leq 0$. Since the functions are piecewise linear, $s_2(2) = 0.5(s_2(1) + s_2(3)) \leq 0$ holds and therefore also $u_2(1) + u_2(3) \leq 0$. Because u_2 must be non-negative, this yields $u_2(1) = u_2(3) = 0$. Similarly, $u_1(1) = u_1(5) = 0$ follows for the upper bound on level 1. Hence, no steps with positive step sizes are possible in this interval. We will now discuss two modifications that stops the "spread of activeness".

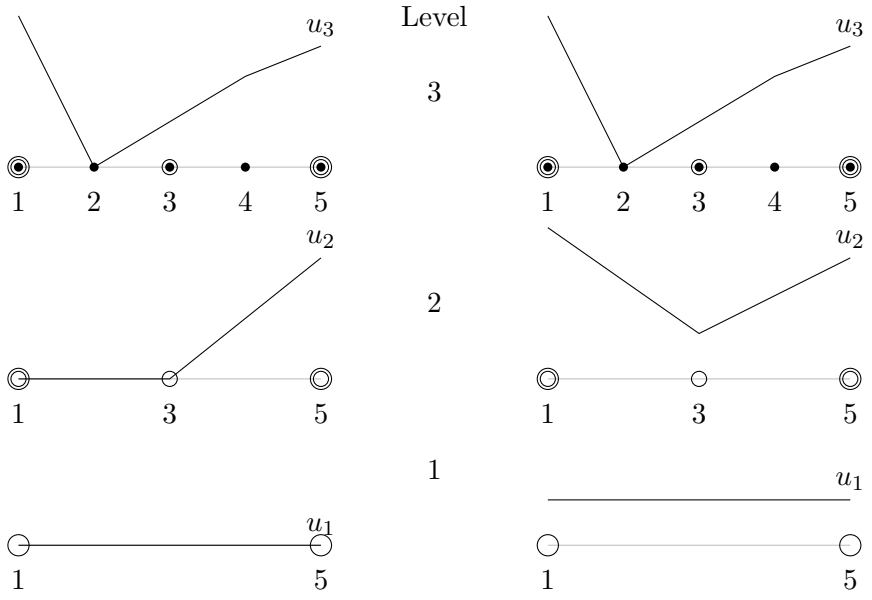

Figure 4.4.: Upper bounds with and without active set strategy

Truncated basis methods

The *truncated basis method* was first presented by Kornhuber in [Kor94] in the context of monotone multigrid methods (see also [Kor97]). The idea is to truncate each coarse grid basis function such that it is zero at active fine-grid points.

Assume $\mathcal{V}_j \subset \mathcal{V}_i$ for all i, j with $j \in N(i)$. As usual in this setting, the identity is supposed to be used as prolongation. Each coarse grid basis function ϕ_j^m can then be written as a linear combination of fine grid nodal basis functions, i.e., we have

$$\phi_j^m(x) = \sum_{l=1}^{n_i} p^{lm} \phi_i^l(x) \tag{4.42}$$

where the p^{lm} are the entries of the prolongation matrix $\tilde{P}_j^i$. Let the current iterate be $v_{i,k}$ and denote the set of active indices by $\mathcal{A}_{i,k} \subset \{1, \ldots, n_i\}$. The truncated coarse grid functions in iteration k are now defined by setting $p^{lm} = 0$ in (4.42) for each $l \in \mathcal{A}_{i,k}$ and $m = 1, \ldots, n_j$. This can be written as follows: Let $\tilde{N}_{i,k} = \mathrm{Diag}(d_{i,k})$ a diagonal matrix where the elements of $d_{i,k}$ are

$$d_{i,k}^l = \begin{cases} 0 & \text{if } l \in \mathcal{A}_{i,k}, \\ 1 & \text{otherwise.} \end{cases}$$

The truncated basis functions $\hat{\phi}_j^m$ are now defined by

$$(\hat{\phi}_j^1, \hat{\phi}_j^2, \ldots, \hat{\phi}_j^{n_j})^T = (\tilde{P}_j^i)^T \tilde{N}_{i,k} (\phi_i^1, \phi_i^2, \ldots, \phi_i^{n_i})^T. \tag{4.43}$$

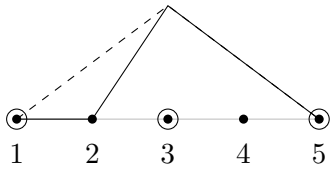

Figure 4.5.: Example of a truncated basis function (solid) in comparison to the usual nodal basis function (dashed) if one node is active

We set $\hat{\mathcal{V}}_j = \operatorname{span}(\{\hat{\phi}_j^k\}_{k=1}^{n_j})$. The important property of this space is that the prolongation of a step $\hat{s}_j \in \hat{\mathcal{V}}_j$ is zero at active components. Therefore, the active points are irrelevant when determining the lower-level bounds. By (4.43) it also follows that $\tilde{T}_j^i = \tilde{N}_{i,k}\tilde{P}_j^i$ is the transfer operator for the coefficient vectors. Note that the prolongation is not changed, only its matrix representation with regard to the truncated generating system $\{\hat{\phi}_j^k\}_k$, which need not be a basis anymore.

The truncated space $\hat{\mathcal{V}}_j$ is a subset of $\mathcal{V}_i$, but it is in general different from $\mathcal{V}_j$ and even $\hat{\mathcal{V}}_j \not\subset \mathcal{V}_j$ holds in most cases. Therefore, we cannot directly use the functions $f_j(x_j, \cdot)$ since they are defined on $\mathcal{V}_j$. In many cases, however, it is possible to define the functions also on the truncated space $\hat{\mathcal{V}}_j$, but the calculation of the function values for a given coefficient vector can be more expensive. This is the case, for instance, in a finite element setting, since the truncated functions $\hat{\phi}_j^l$ are in general not piecewise linear on each coarse grid triangle anymore. Therefore, often more sophisticated and expensive quadrature formulas must be used to obtain the same accuracy. Another problem is that already existing finite element software must be adapted to support these basis functions.

One example where this method is applicable without further modifications is if $f_j \equiv 0$ for $j \neq r$ and the second-order corrected model (2.17) is used: Let $v_{i,k}$ be the current iterate and $g_{i,k} = \nabla h_i(v_{i,k})$ and $H_{i,k} \colon \mathcal{V}_i \to \mathcal{V}_i$ be the gradient and the Hessian approximation of h_i at $v_{i,k}$. On level j, $j \in N(i)$, we obtain a simple quadratic model $h_j \colon \hat{\mathcal{V}}_j \to \mathbb{R}$,

$$h_j(s_j) = (\!(s_j, g_{i,k})\!)_i + \frac{1}{2}(\!(s_j, H_{i,k}s_j)\!)_i.$$

In particular, the representation with regard to the coefficient vectors of $\hat{\mathcal{V}}_j$ is given by:

$$\tilde{h}_j(\tilde{s}_j) = \tilde{s}_j^T(\tilde{T}_j^i)^T \underset{\sim}{g}_{i,k} + \frac{1}{2}\tilde{s}_j^T(\tilde{T}_j^i)^T \underset{\approx}{H}_{i,k}\tilde{T}_j^i\tilde{s}_j.$$

Therefore, we only have to calculate the vector $\underset{\sim}{g}_j = (\tilde{T}_j^i)^T \underset{\sim}{g}_{i,k}$ and the matrix $\underset{\approx}{H}_j = (\tilde{T}_j^i)^T \underset{\approx}{H}_{i,k}\tilde{T}_j^i$ when entering level j.

Remark 4.12 It is obvious that if we use the lower-level sets obtained by the truncated basis method to calculate the stationarity measure χ_i^{ML}, we lose the continuity since the active set changes discontinuously, which influence both, the lower-level sets and the projection operator Q_i^j.

Active set strategy

Our second approach is also applicable if the lower-level functions f_i cannot be changed as necessary for the truncated basis ansatz or the spaces $\mathcal{V}_i$ are not nested as, for instance, in Example 2.2. The idea is simple: Instead of changing the lower-level basis, we just use a modified prolongation operator, which sets the coefficients of active indices to zero. This is achieved by the transfer operator used in the truncated basis method. Instead of just considering active points, we use a larger set of ε-active points: Let $g_{i,k} = \nabla h_i(v_{i,k})$. Define

$$\begin{aligned} \mathcal{A}^\varepsilon_{i,k} := \mathcal{A}^-_{i,k} \cup \mathcal{A}^+_{i,k},\ \mathcal{A}^-_{i,k} &:= \{0 \le j \le n_i |\ \tilde{v}^j_{i,k} - \tilde{l}^j_i \le \varepsilon^A_{i,k} \text{ and } -\tilde{g}^j_{i,k} \le 0\}, \\ \mathcal{A}^+_{i,k} &:= \{0 \le j \le n_i |\ \tilde{u}^j_i - \tilde{v}^j_{i,k} \le \varepsilon^A_{i,k} \text{ and } -\tilde{g}^j_{i,k} \ge 0\}. \end{aligned} \tag{4.44}$$

The motivation behind this definition is that for a suitable choice of $\varepsilon^A_{i,k}$, e.g., $\varepsilon^A_{i,k} = \varepsilon^A \|\tilde{g}_{i,k}\|_\infty$ with $\varepsilon^A \in (0,1)$, we expect that $\mathcal{A}^\varepsilon_{i,k}$ is a better approximation of the set of indices that are active in the solution than the indices in $\mathcal{A}_{i,k}$. When we enter Step 2 in iteration (i,k) of Algorithm 2.1 we define the transfer operator $T^{i,k}_j \colon \mathcal{V}_j \to \mathcal{V}_i$ that is given by its matrix representation

$$\tilde{T}^{i,k}_j = \tilde{N}_{i,k} \tilde{P}^i_j. \tag{4.45}$$

The matrix $\tilde{N}_{i,k} = \mathrm{Diag}(d_{i,k})$ is diagonal and the entries of the vector $d_{i,k} \in \{0,1\}^{n_i}$ are given by

$$d^l_{i,k} = \begin{cases} 0 & \text{if } l \in \mathcal{A}^\varepsilon_{i,k}, \\ 1 & \text{if } l \notin \mathcal{A}^\varepsilon_{i,k}. \end{cases}$$

This new transfer operator first prolongates the step using the standard prolongation and afterwards sets all active indices to zero. This shows that active indices on the fine level do not limit steps on the lower level. Now the transfer operators $T^{i,k}_j$ are used in Algorithm 2.1 instead of the prolongations P^j_i. In particular, (4.34) becomes

$$0 \in C_j \quad \text{and} \quad \tilde{l}_i \le \tilde{v}_{i,k} + \tilde{T}^{i,k}_j \tilde{s}_j \le \tilde{u}_i \text{ for all } s_j \in C_j. \tag{4.46}$$

A lower bound that satisfies (4.46) in our example using $\tilde{T}^2_1 = \mathrm{Diag}((1,0,1,1,1))\tilde{P}^2_1$ is shown on the right side of Figure 4.4. We see that the new bound is far less restrictive, the "activeness" does not spread. In general, assumption (4.46) allows larger lower-level sets than (4.34), i.e., each closed and convex set C_j satisfying (4.34) also satisfies (4.46): Consider a box C_i with bounds l_i and u_i. For a step s_j with $v_{i,k} + P^i_j s_j \in C_i$, we obtain for the coefficients

$$(\tilde{v}_{i,k} + \tilde{T}^{i,k}_j \tilde{s}_j)^m = \tilde{v}^m_{i,k} + d^m_{i,k} (\tilde{P}^i_j \tilde{s}_j)^m = \begin{cases} \tilde{v}^m_{i,k} & \text{if } m \in \mathcal{A}^\varepsilon_{i,k}, \\ (\tilde{v}_{i,k} + \tilde{P}^i_j s_j)^m & \text{if } m \notin \mathcal{A}^\varepsilon_{i,k}, \end{cases} \in [\tilde{l}^m_i, \tilde{u}^m_i]$$

for all $m = 1, \dots, n_i$ and therefore $v_{i,k} + T^j_i s_j \in C_i$ holds.

One downside of this approach is that contrary to the truncated basis method, there could be steps $s_j \neq 0$ with $T^{i,k}_j s_j = 0$ for which $h_j(s_j) \neq 0$. This results in a poor performance of the lower-level steps and the trust regions become small. This is not surprising: Consider the case

where all fine-grid bounds are active, then by (4.46) we are allowed to use $C_j = \mathcal{V}_j$, which is obviously not a good approximation of C_i on the coarse grid. One way to prevent this is to constraint the lower-level set C_j besides (4.46) so that no direction $s_j \neq 0$ is included that satisfies $T_j^{i,k} s_j = 0$. For the case of nodal basis functions where the set of lower-level nodes $\mathcal{N}_j$ is a subset of $\mathcal{N}_i$, this can be achieved by setting $\tilde{l}_j^m = \tilde{u}_j^m = 0$ for indices m for which $(T_j^i \phi_j^m)(x_j^m) = 0$, $x_j^m \in \mathcal{N}_j$, holds.

Additional attention must be paid to the choice of the trust-region norm since assumption (2.20) must hold for the new transfer operators. In the case of linear finite elements, the $L^p(\Omega)$-norms for $p = 1, \ldots, \infty$ obviously satisfy the assumption. The same is true if $\mathcal{V}_i = \mathbb{R}^{n_i}$, $\tilde{P}_j^i \geq 0$, and the trust-region norms are similar to $\|\cdot\|_p$, $p \in \{1, 2, \infty\}$ (properly scaled if necessary). However, if the trust-region norm is $\|\cdot\|_{H^1(\Omega)}$, the assumption is in general not satisfied, since setting the step to zero at singular nodes can introduce oscillations, which increase its norm. This happens only at the boundary of the active set which in most cases consists just of a small number of nodes and hence the effect of the additional oscillations is negligible. In our numerical implementation we added a watchdog, which refuses a lower-level step that violates the trust-region condition on the originating level too much. In this case, the multilevel iteration is repeated using the standard prolongation. However, in our numerical tests this never happened.

5. Applications

In this chapter we analyze concrete examples and show how the different parts of the trust-region algorithm can be chosen, such that the assumptions necessary for global convergence of the algorithm are satisfied.

Most assumptions we have made so far were concerned with the function spaces and their discretization. We have showed that in the case that we have suitable discretizations of $H^1(\Omega)$, most of them are satisfied. Hence, we concentrate in this chapter on the assumptions that are related to the functions f_r and the lower-level models h_i. Let us recall them:

(H1) The function f_r and all lower-level models h_i, $i = 1, \dots, r-1$, are continuously differentiable. Furthermore, they are twice Gâteaux differentiable and the mappings $v_r \mapsto f_r''(v_r)[d_r, d_r]$, $d_r \in \mathcal{V}_r$, and $v_i \mapsto h_i''(v_i)[d_i, d_i]$, $d_i \in \mathcal{V}_i$, are continuous.

(H2) There exists a level-independent constant β_1 such that

$$\left|\left\langle (H_{i,k} - h_i''(v_{i,k} + ts_{i,k}))s_{i,k}, s_{i,k}\right\rangle\right| \leq 2\beta_1 \|s_{i,k}\|_i^2$$

holds for all $t \in [0,1]$, iterations $v_{i,k} \in \mathcal{V}_i$ and steps $s_{i,k}$ generated by Algorithm 2.1.

(H3) There exists a level-independent constant β_2 such that

$$\left|\left\langle (h_j''(tv_j) - (P_j^i)^* h_i''(v_{i,k} + tP_j^i v_j)P_j^i)v_j, v_j\right\rangle\right| \leq 2\beta_2 \|v_j\|_j^2$$

is satisfied for all $t \in [0,1]$, multilevel iterations $v_{i,k} \in \mathcal{V}_i$ and steps v_j generated by Algorithm 2.1. Here, h_j is the lower-level model of h_i at $v_{i,k}$.

(H4) The following estimate holds for all Hessian-approximations $H_{i,k}$ with a level-independent constant C_H:

$$\langle v_i, H_{i,k} u_i\rangle \leq C_H \lambda_i^{\max} \|u_i\| \|v_i\| \quad \text{for all } u_i, v_i \in \mathcal{V}_i.$$

(H5) f_r' is uniformly continuous on a set $\mathcal{S} \subset C_r$ that contains the sequence of iterates $(v_{r,k})_{k\in\mathbb{N}}$, i.e. for all $\varepsilon > 0$ there is a $\delta > 0$ such that

$$\|f_r'(v_r) - f_r'(u_r)\|_{\mathcal{V}_r^*} \leq \varepsilon \text{ for all } v_r, u_r \in \mathcal{S} \text{ with } \|v_r - u_r\|_{\mathcal{V}_r} \leq \delta.$$

Assumption (H5) is essential to show the uniform continuity of the stationarity measures, which is needed to show the strong convergence in Theorem 2.2.

In the examples that we are considering in this chapter, we assume an underlying infinite dimensional problem with an objective function $f\colon \mathcal{V} \to \mathbb{R}$. Furthermore, we suppose that the level hierarchy consists of the same problem considered on nested and finite dimensional subspaces with increasing degrees of freedom. More precisely, we assume the following setting:

Assumption 5.1 *1. The spaces $\mathcal{V}_i$ are finite dimensional, $\mathcal{V}_i \subset \mathcal{V}_{i+1}$ for $i = 1, \dots, r-1$ and $\mathcal{V}_r \subset \mathcal{V}$ holds.*

2. The identity is used as prolongation.

3. $\mathcal{W}_i = \mathcal{V}_i$ and $f_i(x_i, v_i) := f(x_i + v_i)$.

4. The first-order corrected model (2.15) are used as lower-level model.

5. The Hessian approximation $H_{i,k}$ in the quadratic models $q_{i,k}$ is the exact Hessian of h_i at $v_{i,k}$.

In this case, the assumptions (H1)–(H4) can be simplified:

(H1') The function f is continuously differentiable. Furthermore, it is twice Gâteaux differentiable and the mappings $v \mapsto f''(v)[d, d]$ are continuous for all $d \in \mathcal{V}$.

(H2') There exists a level-independent constant β_1 such that

$$\left|\langle (f''(x_i + v_{i,k}) - f''(x_i + v_{i,k} + t s_{i,k}))s_{i,k}, s_{i,k}\rangle\right| \le 2\beta_1 \|s_{i,k}\|_i^2$$

holds for all $t \in [0, 1]$, iterations $v_{i,k} \in \mathcal{V}_i$, and steps $s_{i,k}$ generated by Algorithm 2.1.

(H3') There exists a level-independent constant β_3 such that

$$\left|\langle (f''(x_{i-1}) - f''(x_i + v_{i,k}))v_{i-1}, v_{i-1}\rangle\right| \le \beta_3 \|v_{i-1}\|_{i-1}^2 \quad \text{for all } v_{i-1} \in C_{i-1}$$

with $x_{i-1} = R_i^{i-1}(x_i, v_{i,k})$ is satisfied for all iterations $x_i + v_{i,k} \in \mathcal{V}_i$ generated by Algorithm 2.1.

(H4') There is a level independent constant C_H such that for all iterates $x_i + v_{i,k}$ the estimate

$$\langle v_i, f''(x_i + v_{i,k})u_i\rangle \le C_H \lambda_i^{\max} \|u_i\| \, \|v_i\| \quad \text{for all } u_i, v_i \in \mathcal{V}_i.$$

is satisfied.

Lemma 5.1 *Assumptions 5.1 and (H1') - (H4') imply (H1) - (H4).*

Proof Obviously, when the first- or second-order corrected models are used, the differentiability assumptions on f imply (H1). By inserting the definition of the first-order model in (H2), we directly obtain (H2').

To establish (H3), it follows from Remark 2.12 that it is sufficient to show

$$\left|\langle (h_{i-1}''(0) - (P_{i-1}^i)^* h_i''(v_{i,k}) P_{i-1}^i)v_{i-1}, v_{i-1}\rangle\right| \le C\|v_{i-1}\|_{i-1}^2 \quad \text{for all } v_{i-1} \in C_{i-1}. \tag{5.1}$$

Since the identity is used as prolongation, we obtain using (H3')

$$\begin{aligned}
\left|\langle v_{i-1}, (h_{i-1}''(0) - (P_{i-1}^i)^* h_i''(v_{i,k}) P_{i-1}^i)v_{i-1}\rangle\right| &= \left|\langle v_{i-1}, (h_{i-1}''(0) - h_i''(v_{i,k}))v_{i-1}\rangle\right| \\
&= \left|\langle (f''(R_i^{i-1}(x_i, v_{i,k})) - f''(x_i + v_{i,k}))v_{i-1}, v_{i-1}\rangle\right| \\
&\le \beta_3 \|v_{i-1}\|_{i-1}^2,
\end{aligned}$$

which shows the assertion. Finally, (H4) follows directly from $H_{i,k} = h_i''(v_{i,k}) = f''(x_i + v_{i,k})$. □

In this chapter we shall only consider bounded domains Ω with Lipschitz-boundary. Hence, from Theorem A.1 we infer the continuous embedding of $H^1(\Omega)$ into $L^p(\Omega)$ ($H^1(\Omega) \hookrightarrow L^p(\Omega)$) for

$$p \in \begin{cases} [1,\infty) & \text{for } d = 2 \\ [1,p^*] & \text{for } d \geq 3 \end{cases}, \quad p^* := \frac{2d}{d-2}.$$

In these cases, the inequality $\|u\|_{L^p(\Omega)} \leq C\|u\|_{H^1(\Omega)}$ holds true for all $u \in H^1(\Omega)$.

As in the previous chapters, we use a generic constant C which may take different values in the inequalities. It is always assumed to be level-independent and sufficiently large. We sometimes omit the domain Ω in the notation when this information is obvious from the context, i.e., we write L^2 instead of $L^2(\Omega)$.

5.1. Example 1

Let $\Omega \subset \mathbb{R}^d$, $d \leq 3$, be a bounded domain with Lipschitz-boundary. Furthermore, let $a \in L^2(\Omega)$, $b \in L^\infty(\Omega)$ be real valued functions and $A\colon \Omega \to \mathbb{R}^{d\times d}$ be such that $A(x)$ is symmetric for all $x \in \Omega$ and the entries satisfy $a_{ij} \in L^\infty(\Omega)$ for $i,j = 1,\dots,d$. We consider the problem

$$\begin{aligned} &\min_{u\in\mathcal{C}} J_1(u), \quad J_1\colon H^1(\Omega) \to \mathbb{R}, \quad u \mapsto \int_\Omega j_1(x,u(x),\nabla u(x))\,\mathrm{d}x, \\ &\qquad j_1\colon \Omega\times\mathbb{R}\times\mathbb{R}^d \to \mathbb{R},\ j_1(x,u,z) := \frac{1}{2}(z^T A(x) z + b(x)u^2) + a(x)u + \varphi(x,u), \end{aligned} \tag{5.2}$$

where $\varphi\colon \Omega\times\mathbb{R}\to\mathbb{R}$ is measurable in $x \in \Omega$ for each u and twice continuously differentiable in u for almost all $x \in \Omega$. We assume $\mathcal{C} \subset \mathcal{V} := H^1(\Omega)$ to be a nonempty, closed and convex set with the property

$$u \in \mathcal{C},\ \xi \in \mathbb{R},\ u+\xi \in \mathcal{C} \quad \Rightarrow \quad |\xi| < C, \tag{5.3}$$

where the constant C is independent of u and ξ. Examples of feasible sets that satisfies this assumption are subsets of $H^1(\Omega)$ with Dirichlet boundary conditions on (a part of) the boundary. Other examples are sets with pointwise constraints, i.e.,

$$\{v \in H^1(\Omega) \,|\, l^b(x) \leq v(x) \leq u^b(x) \quad \text{for } x \in \Omega \text{ a.e.}\}$$

with $L^\infty(\Omega)$-functions l^b, u^b that satisfy $l^b \leq u^b$ a.e. in Ω. Under this assumptions on the feasible set, the generalized Poincaré's inequality (cf., e.g., [Alt06, Section 6.16]) is satisfied, i.e., it exists a constant $C > 0$ such that

$$\|u\|_{L^2(\Omega)} \leq C(\|\nabla u\|_{L^2(\Omega)} + 1) \quad \text{for all } u \in \mathcal{C}. \tag{5.4}$$

In the following, let $p \geq 2$ be chosen such that $H^1(\Omega) \hookrightarrow L^p(\Omega)$. Then, from the embedding and (5.4) we infer

$$\|u\|_{L^p(\Omega)} \leq C\|u\|_{H^1(\Omega)} \leq C(\|\nabla u\|_{L^2(\Omega)} + 1) \quad \forall u \in \mathcal{C}. \tag{5.5}$$

A, a, b and φ shall be chosen such that J_1 admits a minimizer. We are going to consider the case that there exist $\varepsilon > 0$, $1 \le s < 2$ and $C \ge 0$ such that

$$\begin{aligned} z^T A(x) z &\ge \varepsilon z^T z && \forall x \in \Omega \text{ and } z \in \mathbb{R}^d, \\ b(x) &\ge 0 && \forall x \in \Omega, \\ \varphi(x,u) &\ge -C(|u|^s + 1) && \forall x \in \Omega \text{ and } u \in \mathbb{R}. \end{aligned}$$

Under these assumptions, by using the generalized Poincaré's inequality and Hölder's inequality, we can estimate

$$\begin{aligned} J_1(u) &\ge \int_\Omega \Big(\varepsilon \|\nabla u\|^2 - |a(x)||u| - C(|u|^s + 1) \Big) \,\mathrm{d}x \\ &\ge \varepsilon \|\nabla u\|^2_{L^2(\Omega)} - \|a\|_{L^2(\Omega)} \|u\|_{L^2(\Omega)} - C(\|u\|^s_{L^s(\Omega)} + \operatorname{area}(\Omega)) \\ &\ge \varepsilon \|\nabla u\|^2_{L^2(\Omega)} - C\|a\|_{L^2(\Omega)} \|\nabla u\|_{L^2(\Omega)} - C(\|\nabla u\|^s_{L^2(\Omega)} + \operatorname{area}(\Omega)) - C \\ &\to \infty \quad \text{as } \|\nabla u\|_{L^2(\Omega)} \to \infty. \end{aligned}$$

This shows that J_1 is a coercive function. Together with the weak lower semicontinuity of J_1 (cf. Theorem A.7 and Remark A.4), Theorem A.6 yields the existence of a minimizer u^* of problem (5.2).

To ensure the necessary differentiability of J_1, we further assume the following growth assumptions:

$$|\varphi(x,u)| \le C(g_1(x) + |u|^q), \qquad g_1 \in L^1(\Omega) \tag{5.6a}$$

$$|\varphi_u(x,u)| \le C(g_2(x) + |u|^{q-1}), \qquad g_2 \in L^{q/(q-1)}(\Omega) \tag{5.6b}$$

$$|\varphi_{uu}(x,u)| \le C(g_3(x) + |u|^{q-2}), \qquad g_3 \in L^{q/(q-2)}(\Omega). \tag{5.6c}$$

with $2 \le q \le p$. Here and subsequently, we often denote the partial derivatives by indices as for example $\varphi_u = \partial\varphi/\partial u$. If $\varphi_{uu}(x,u)$ is bounded, we can set $q = 2$ and have $L^{q/(q-2)}(\Omega) = L^\infty(\Omega)$.

The next lemma shows that the functional J_1 satisfies the differentiability assumptions (H1').

Lemma 5.2 *Under the assumptions* (5.6)*, the functional J_1 is twice Gâteaux differentiable on $H^1(\Omega)$. Furthermore, the first derivative is continuous and the operator $u \mapsto J_1''(u)[d,d]$ is continuous for every fixed direction $d \in H^1(\Omega)$.*

Proof We show that the function j_1 satisfies assumptions (A.4). Using the growth condition (5.6a) and Young's inequality, we obtain for almost all $x \in \Omega$:

$$\begin{aligned} |j_1(x,u,z)| &\le \|A(x)\| \|z\|^2 + |b(x)|u^2 + a(x)|u| + C(g_1(x) + |u|^q) \\ &\le \|A\|_{L^\infty(\Omega)^{d\times d}} \|z\|^2 + \|b\|_{L^\infty(\Omega)} u^2 + \frac{1}{2}(a(x)^2 + u^2) + C(g_1(x) + |u|^q). \end{aligned}$$

Since $q \ge 2$ and $a \in L^2(\Omega)$, there exists a constant C and a $L^1(\Omega)$-function $\bar{g}_1$ such that the following bound holds:

$$|j_1(x,u,z)| \le C(\bar{g}_1(x) + \|z\|^2 + |u|^q) \quad \text{a.e. on } \Omega.$$

Similarly, we obtain with $\bar{g}_2$ being a suitable $L^{q/(q-1)}(\Omega)$-function the following estimates for the partial derivatives almost everywhere on Ω:

$$\begin{aligned}\left|\frac{\partial j_1}{\partial u}(x,u,z)\right| &= |b(x)u + a(x) + \varphi_u(x,u)| \le C(\bar{g}_2(x) + |u|^{q-1}),\\ \left\|\frac{\partial j_1}{\partial z}(x,u,z)\right\| &= \|A(x)z\| \le C\|z\|.\end{aligned}$$

Together with Remark A.2, Theorem A.4 shows that J_1 is continuously differentiable.

In the same way, one estimates the second-order partial derivatives of j_1 and uses Theorem A.5 in consideration of Remark A.3 to obtain the second-order differentiability. □

The next lemma shows that J_1' satisfies (H5) under suitable assumptions on φ_u.

Lemma 5.3 *Let Φ_u with $\Phi_u(u)(x) := \varphi_u(x,u(x))$ be uniformly continuous as mapping from $L^q(\Omega)$ to $L^{q/(q-1)}(\Omega)$, $2 \le q \le p$, on a set $\mathcal{S} \subset H^1(\Omega)$, i.e., for all $\varepsilon_{\Phi_u} > 0$ exists a $\delta_{\Phi_u} > 0$ such that*

$$\|\Phi_u(u) - \Phi_u(v)\|_{L^{q/(q-1)}(\Omega)} \le \varepsilon_{\Phi_u} \quad \text{for all } u,v \in \mathcal{S} \text{ with } \|u-v\|_{L^q(\Omega)} \le \delta_{\Phi_u}.$$

Then J_1' is uniformly continuous on $\mathcal{S}$.

Proof Let $\varepsilon > 0$. We set $\varepsilon_{\Phi_u} = \varepsilon/(2C)$, where C must be chosen large enough such that $\|w\|_{L^q(\Omega)} \le C\|w\|_{H^1(\Omega)}$ for all $w \in H^1(\Omega)$ hold. The existence of such a constant is assured by the embedding $H^1(\Omega) \hookrightarrow L^q(\Omega)$. We denote by δ_{Φ_u} the corresponding δ of the uniform continuity of Φ_u. The definition of the dual norm, Hölder's inequality and the embedding of $H^1(\Omega)$ into $L^q(\Omega)$ yield

$$\begin{aligned}\|J_1'(u) - J_1'(v)\|_{\mathcal{V}^*} &= \sup_{\|d\|_{\mathcal{V}}=1} \langle J_1'(u) - J_1'(v), d\rangle\\ &= \int_\Omega \Big(\nabla d^T A(x)(\nabla u - \nabla v) + b(x)d(u-v) + d(\varphi_u(x,u) - \varphi_u(x,v))\Big)\,\mathrm{d}x\\ &\le \|A\|_{L^{\infty}{}^{d\times d}}\|\nabla u - \nabla v\|_{L^2}\|\nabla d\|_{L^2} + \|b\|_{L^\infty}\|d\|_{L^2}\|u-v\|_{L^2}\\ &\quad + \|d\|_{L^q}\|\Phi_u(u) - \Phi_u(v)\|_{L^{q/(q-1)}}\\ &\le \|A\|_{L^{\infty}{}^{d\times d}}\|\nabla u - \nabla v\|_{L^2} + \|b\|_{L^\infty}\|u-v\|_{L^2} + C\|\Phi_u(u) - \Phi_u(v)\|_{L^{q/(q-1)}}.\end{aligned}$$

Now let $u, v \in \mathcal{S}$ arbitrary with

$$\|u-v\|_{\mathcal{V}} \le \delta := \min\Big\{\delta_{\Phi_u}, \frac{\varepsilon}{2}(\|A\|_{L^{\infty}{}^{d\times d}} + \|b\|_{L^\infty})^{-1}\Big\}.$$

Then we have

$$\|J_1'(u) - J_1'(v)\|_{\mathcal{V}^*} \le (\|A\|_{L^{\infty}{}^{d\times d}} + \|b\|_{L^\infty})\delta + C\varepsilon_{\Phi_u} \le \varepsilon,$$

which shows the assertion. □

Remark 5.1 The uniform continuity of Φ_u is weaker than demanding that the function $\varphi_u(x,\cdot)$ is uniformly continuous for almost all $x \in \Omega$. As an example consider the function $\varphi(x,u) := u^q/q$, $q > 2$, for which $\varphi_u(x,\cdot)$ is not uniformly continuous on $\mathbb{R}$. But if we assume that $\|u\|_{L^q} \leq C$ for all $u \in \mathcal{S}$, one can show, using the inequality $(u^{q-1} - v^{q-1}) \leq (q-1)(|u|^{q-2} + |v|^{q-2})|u-v|$, that there exists a $C > 0$ such that

$$\|\Phi_u(u) - \Phi_u(v)\|_{L^{q/(q-1)}} = \|u^{q-1} - v^{q-1}\|_{L^{q/(q-1)}} \leq C\|u-v\|_{L^q},$$

which shows the uniform continuity of Φ_u in the sense of the previous lemma.

We continue to verify the remaining assumptions (H2')–(H4') for which we need the second-order directional derivatives, which are given by

$$J_1''(u)[d_1,d_2] = \int_\Omega \Big(\nabla d_2{}^T A(x)\nabla d_1 + \big(b(x) + \varphi_{uu}(x,u(x))\big)d_1 d_2\Big)\,\mathrm{d}x. \tag{5.7}$$

We now show that the induced bilinear form is bounded on $H^1(\Omega)$ for all iterates of Algorithm 2.1, i.e., there is a constant C_H independent of i, k and t such that

$$J_1''(x_i + v_{i,k} + ts_{i,k})[d_1,d_2] \leq C_H\|d_1\|_{H^1(\Omega)}\|d_2\|_{H^1(\Omega)}.$$

To show this, we define $\Phi_{uu}(u)(x) := \varphi_{uu}(x,u(x))$ and assume that it is bounded in $L^{q/(q-2)}(\Omega)$ for all iterates generated by Algorithm 2.1, i.e., there exists a constant C_Φ such that

$$\|\Phi_{uu}(x_i + v_{i,k} + ts_{i,k})\|_{L^{q/(q-2)}(\Omega)} \leq C_\Phi \text{ for all } t \in [0,1], i = 1,\dots,r \text{ and all iterations } k = 1,\dots. \tag{5.8}$$

Now Hölder's inequality and the embedding $H^1(\Omega) \hookrightarrow L^q(\Omega)$ yield for $t \in [0,1]$

$$\begin{aligned} J_1''(x_i + v_{i,k} + ts_{i,k})[d_1,d_2] &\leq \|A\|_{L^{\infty}{}^{d\times d}}\|\nabla d_1\|_{L^2}\|\nabla d_2\|_{L^2} + \|b\|_{L^\infty}\|d_1\|_{L^2}\|d_2\|_{L^2} \\ &\quad + \|\Phi_{uu}(x_i + v_{i,k} + ts_{i,k})\|_{L^{q/(q-2)}}\|d_1\|_{L^q}\|d_2\|_{L^q} \\ &\leq C_H\|d_1\|_{H^1}\|d_2\|_{H^1}, \end{aligned} \tag{5.9}$$

where $C_H = \max\{\|A\|_{L^{\infty}{}^{d\times d}}, \|b\|_{L^\infty}\} + CC_\Phi$. With the definition of $\lambda_i^{\max}$, (3.12), follows (H4').

In the next step, we want to determine how to choose a trust-region norm such that (H2') and (H3') are satisfied. It follows immediately from (5.9) that both assumptions hold with $\beta_1 = \beta_3 = 2C_H$ if $\|\cdot\|_{H^1(\Omega)}$ is chosen as trust-region norm. We will now show that it is also possible to choose a weaker norm for the trust region in this example. Using Hölder's inequality it follows from (5.7):

$$\begin{aligned} &\Big|(J_1''(x_i + v_{i,k}) - J_1''(x_i + v_{i,k} + ts_{i,k}))[s_{i,k},s_{i,k}]\Big| \\ &\qquad \leq \int_\Omega \Big|(\varphi_{uu}(x,x_i + v_{i,k}) - \varphi_{uu}(x,x_i + v_{i,k} + ts_{i,k}))s_{i,k}^2\Big|\,\mathrm{d}x \\ &\qquad \leq \|\Phi_{uu}(x_i + v_{i,k}) - \Phi_{uu}(x_i + v_{i,k} + ts_{i,k})\|_{L^{q/(q-2)}(\Omega)}\|s_{i,k}\|_{L^q(\Omega)}^2. \end{aligned}$$

Hence, if (5.8) holds, then also (H2') and (H3') for $\|\cdot\|_i = \|\cdot\|_{L^q(\Omega)}$, $i = 1,\dots,r$.

The requirement that Φ_{uu} is bounded on $\mathcal{C}$, which implies (5.8), is strong, in particular in the unconstrained case. We will now show that (5.8) is satisfied in our setting for a slightly changed algorithm if the restriction operators are stable with respect to the trust-region norm. To this end, we use the following lemma:

Lemma 5.4 *Let Assumption 5.1 be satisfied. Suppose that $\hat{x}_r \in C_r$ is a point such that its sublevel set,*

$$L^-_{\hat{x}_r}(f) := \{x_r \in \mathcal{V}_r \cap C_r \mid f_r(x_r) \leq f_r(\hat{x}_r)\},$$

is bounded with respect to a norm $\|\cdot\|_\#$, which satisfies $\|v_i\|_\# \leq \|v_i\|_i$ for all $v_i \in \mathcal{V}_i$ and $i = 1, \ldots, r$. Assume that there are linear restriction operators $R_i \colon \mathcal{V}_r \to \mathcal{V}_i$ for $i = 1, \ldots, r-1$ such that

$$R^i_{i+1}(x_{i+1}, v_{i+1}) = R_i(x_{i+1} + v_{i+1})$$

and the estimate

$$\|R_i(R_{i+1}(\cdots(R_j(v_{j+1})))\|_\# \leq C_R \|v_{j+1}\|_\# \text{ for all } i \leq j < r \tag{5.10}$$

holds with a level-independent constant C_R. Then all points $x_i + v_{i,k} + ts_{i,k}$, $t \in [0,1]$, that are generated by Algorithm 2.1, applied to f_r with start point $\hat{x}_r$, are bounded indepedently of the level with respect to $\|\cdot\|_\#$ if there exists a maximum trust-region radius $\Delta_{r,\max} > 0$ such that $\Delta_{r,k} \leq \Delta_{r,\max}$ for all k holds.

Proof Since $\|s_{i,k}\|_\# \leq \|s_{i,k}\|_i \leq \Delta_{i,k} \leq \Delta_{r,\max}$ holds, it is enough to show that the iterates $x_i + v_{i,k}$ are bounded.

Algorithm 2.1 is a descent method and thus all iterates $x_r + v_{r,k} = v_{r,k}$ on level r stay inside the sublevel set. Since the sublevel set is bounded, there is a constant B such that $\|v_{r,k}\|_\# \leq B$ holds.

In contrast to $h_r = f_r$, the lower-level models h_i on levels $i < r$ can be unbounded from below. However, due to the trust-region management, the iterates on the lower levels also stay bounded: Let iteration (i, k_i) be generated by iteration $(i+1, k_{i+1})$ which itself was generated by iteration $(i+2, k_{i+2})$ and so on till an iteration (r, k_r). From Corollary 2.1, we infer

$$\|x_i + v_{i,k_i}\|_\# \leq \|x_i\|_\# + \|v_{i,k_i}\|_i \leq \|x_i\|_\# + C_\mathcal{P} \Delta_{i,0}.$$

Without loss of generality, we require $C_\mathcal{P} \geq C_R \geq 1$. We recall that by construction $x_j = R_j(x_{j+1} + v_{j+1,k_{j+1}})$ for all $j = 1, \ldots, r-1$ is satisfied. Hence, using (5.10) we can estimate

$$\begin{aligned}\|x_i\|_\# &= \|R_i(x_{i+1} + v_{i+1,k_{i+1}})\|_\# \leq \|R_i R_{i+1}(x_{i+2} + v_{i+2,k_{i+2}})\|_\# + \|R_i v_{i+1,k_{i+1}}\|_\# \\ &\leq \ldots \leq \sum_{j=i}^{r-1} \|R_i \cdots R_j v_{j+1,k_{j+1}}\|_\# \leq C_R \sum_{j=i+1}^{r} \|v_{j,k_j}\|_\#\end{aligned}$$

This yields

$$\|x_i + v_{i,k_i}\|_\# \leq C_R \sum_{j=i+1}^{r} \|v_{j,k_j}\|_\# + C_\mathcal{P} \Delta_{i,0} \leq C_\mathcal{P} \sum_{j=i+1}^{r} \|v_{j,k_j}\|_\# + C_\mathcal{P} \Delta_{i+1,k_{i+1}}.$$

From the second part of the trust-region update rule (2.34), it follows

$$C_\mathcal{P} \Delta_{j,k_j} \leq C_\mathcal{P} \Delta_{j,0} - C_\mathcal{P} \|v_{j,k_j}\|_j \leq C_\mathcal{P} \Delta_{j,0} - C_\mathcal{P} \|v_{j,k_j}\|_\# \quad \text{for all } i+1 \leq j < r$$

and hence

$$\|x_i + v_{i,k_i}\|_{\#} \leq C_{\mathcal{P}} \sum_{j=i+2}^{r} \|v_{j,k_j}\|_{\#} + C_{\mathcal{P}} \Delta_{i+2,k_{i+2}} \leq \ldots \leq C_{\mathcal{P}} \|v_{r,k_r}\|_{\#} + C_{\mathcal{P}} \Delta_{r,k_r} \leq C_{\mathcal{P}}(B + \Delta_{r,\max}).$$

This completes the proof. □

Remark 5.2 The additional assumption, $\Delta_{r,k} \leq \Delta_{r,\max}$, can be guaranteed by changing the update rule (2.34) in Algorithm 2.1 to

$$\Delta_{i,k+1} = \begin{cases} \min\{\Delta_{i,k}^{+}, \Delta_{i,0} - \|v_{i,k+1}\|_i\} & \text{if } i < r, \\ \min\{\Delta_{i,k}^{+}, \Delta_{r,\max}\} & \text{if } i = r. \end{cases}$$

The global convergence properties of the algorithm are not affected by this change.

Now let $\|\cdot\|_i = \|\cdot\|_{L^q(\Omega)}$ for $i = 1, \ldots, r$. From the coercivity of J_1 and (5.5) follows the boundedness of all sublevel sets of J_1 with respect to the $L^q(\Omega)$-norm. Set $\|\cdot\|_{\#} = \|\cdot\|_{L^q(\Omega)}$ and let the assumptions of the preceding lemma on the restrictions hold.[1] Let $q \neq 2$ and let $x_i + v_{i,k} + ts_{i,k}$, $t \in [0,1]$, be an arbitrary iterate. Then from the growth condition (5.6) on φ_{uu}, it follows

$$\begin{aligned} \|\Phi_{uu}(x_i + v_{i,k} + ts_{i,k})\|_{L^{q/(q-2)}} &\leq \Big(\int_\Omega |\varphi_{uu}(x, x_i + v_{i,k} + ts_{i,k})|^{q/(q-2)} \, \mathrm{d}x \Big)^{(q-2)/q} \\ &\leq C \left[\|g_3\|_{L^{q/(q-2)}} + \Big(\int_\Omega |x_i + v_{i,k} + ts_{i,k}|^q \, \mathrm{d}x \Big)^{(q-2)/q} \right] \\ &\leq C \big[\|g_3\|_{L^{q/(q-2)}} + \|x_i + v_{i,k} + ts_{i,k}\|_{L^q}^{q-2} \big]. \end{aligned}$$

Lemma 5.4 shows that the last expression is bounded independently of i and k and hence there exists a constant C_Φ such that (5.8) is satisfied. In the case $q = 2$, (5.8) becomes

$$\|\Phi_{uu}(x_i + v_{i,k} + ts_{i,k})\|_{L^\infty(\Omega)} \leq C_\Phi$$

and follows directly from the growth condition on φ_{uu}.

By a similar argumentation using (5.5), it can be easily seen that we obtain the same result for $\|\cdot\|_{\#} = \|\cdot\|_i = \|\cdot\|_{H^1(\Omega)}$.

5.2. A quasi-interpolation restriction operator

Before we consider other application classes, we present in this section two different restriction operators that satisfy the assumptions of Lemma 5.4.

It is easy to see that the simple nodal interpolation operator (*injection*) generally do not satisfy condition (5.10) if $\|\cdot\|_{L^q(\Omega)}$ or $\|\cdot\|_{H^1(\Omega)}$ is chosen as trust-region norm. One possible choice for a restriction operator that satisfies (5.10) both for $\|\cdot\|_{\#} = \|\cdot\|_{H^1(\Omega)}$ and $\|\cdot\|_{\#} = \|\cdot\|_{L^q(\Omega)}$ is the

[1] For a restriction operator that satisfies (5.10) see Section 5.2.

$L^2(\Omega)$-projection Q_i. To see this, we first note that $Q_i(Q_j)v = Q_i v$ for all $v \in L^2(\Omega)$ and $i \leq j$, which follows directly from the orthogonality (cf. (3.2)):

$$\begin{aligned}\|Q_i(Q_j v) - Q_i v\|^2_{L^2(\Omega)} &= (Q_i(Q_j v - v), Q_i(Q_j v - v))_{L^2(\Omega)} = (Q_i(Q_j v - v), Q_j v - v)_{L^2(\Omega)} \\ &= (Q_i(Q_j v - v), Q_j v)_{L^2(\Omega)} - (Q_i(Q_j v - v), v)_{L^2(\Omega)} = 0.\end{aligned}$$

Therefore, it is enough to show the stability, i.e., that $\|Q_i v\|_{\#} \leq C_R \|v\|_{\#}$ holds. For the H^1-norm this is a well-known result (cf. for instance [BX91, Thm. 3.4]). In [DDW75] a stability result for the L^p-norms was proven, more precisely, it was shown that

$$\|Q_i v\|_{L^p(\Omega)} \leq C^{|1-2/p|} \|v\|_{L^p(\Omega)} \quad \text{for all } v \in L^p(\Omega) \text{ and } 1 \leq p \leq \infty$$

holds for a large number of finite element spaces over quasi-uniform grids, for instance in the setting of Example 3.1. The constant C is level-independent and does not depend on p.

We now present another restriction operator, which can be numerically evaluated cheaper than the L^2-projection. For this, we require that each space $\mathcal{V}_i$ is equipped with a basis $\{\phi_i^j\}_{j=1,\dots,n_i} \subset W^{1,\infty}(\Omega)$ that satisfies the following assumptions:

1. $\phi_i^j \geq 0$ for all $j = 1, \dots, n_i$,
2. $\|\phi_i^j\|_{L^\infty(\Omega)} = 1$,
3. $0 < \theta_i \leq 1$ almost everywhere in Ω where $\theta_i := \sum_{j=1}^{n_i} \phi_i^j$.

A typical example of a basis that satisfies these conditions is the nodal basis presented in Example 3.1.

We define quasi-interpolation operators I_i by

$$I_i \colon L^1(\Omega) \to \mathcal{V}_i, \quad I_i(u) := \sum_{j=1}^{n_i} \pi_i^j(u) \phi_i^j \tag{5.11}$$

where

$$\pi_i^j(u) := \frac{(u, \phi_i^j)}{(\phi_i^j, 1)} = \int_\Omega u \phi_i^j \,\mathrm{d}x \Big(\int_\Omega \phi_i^j \,\mathrm{d}x \Big)^{-1}.$$

These types of quasi-interpolation operators were also be considered in [BPV00] and [Car99]. In the latter they analysed a slightly different interpolation where the coefficients are defined by $\hat{\pi}_i^j(u) := \int_\Omega u \phi_i^j / \theta_i \,\mathrm{d}x \Big(\int_\Omega \phi_i^j \,\mathrm{d}x \Big)^{-1}$. In the setting of Example 3.1 the operators differ only at nodes near the Dirichlet boundary.

Remark 5.3 We emphasize that in comparison to the L^2-projector the evaluation of the quasi-interpolant is inexpensive since no linear system involving the mass-matrix has to be solved.

We first show the stability with respect to the $L^p(\Omega)$-norms.

Lemma 5.5 *The interpolations defined by* (5.11) *are continuous and linear operators from* $L^p(\Omega)$ *to* $L^p(\Omega)$ *for* $1 \leq p \leq \infty$. *In particular, for* $u \in L^p(\Omega)$ *it holds:*

$$\|I_i u\|_{L^p(\Omega)} \leq \|u\|_{L^p(\Omega)}. \tag{5.12}$$

Proof Obviously, the interpolation is a linear operator and well-defined since the basis functions satisfy $\phi_i \in L^\infty(\Omega)$ for $i = 1, \ldots, n_i$. We show (5.12) for $p = 1$ and $p = \infty$. From this, the assertion then follows by the Riesz-Thorin interpolation theorem (cf. for instance [Wer07, Thm. II.4.2]).

$\mathbf{p = 1}$: Using $\phi_i^j \geq 0$ and $\sum_{j=1}^{n_i} \phi_i^j \equiv 1$ a.e. on Ω, we obtain

$$\|I_i u\|_{L^1(\Omega)} \leq \sum_{j=1}^{n_i} \left|\pi_i^j(u)\right| \int_\Omega |\phi_i^j| \,\mathrm{d}x \leq \sum_{j=1}^{n_i} \int_\Omega |u| \phi_i^j \,\mathrm{d}x \leq \int_\Omega |u| \,\mathrm{d}x = \|u\|_{L^1(\Omega)}. \tag{5.13}$$

$\mathbf{p = \infty}$: Let $u \in L^\infty(\Omega)$. Clearly, $|\pi_i^j(u)| \leq \|u\|_{L^\infty(\Omega)}$ and hence

$$\|I_i u\|_{L^\infty(\Omega)} \leq \|\sum_{j=1}^{n_i} \pi_i^j(u) \phi_i^j\|_{L^\infty(\Omega)} \leq \|u\|_{L^\infty(\Omega)} \|\sum_{j=1}^{n_i} \phi_i^j\|_{L^\infty(\Omega)} \leq \|u\|_{L^\infty(\Omega)}.$$

□

We will now show that this interpolation operator also satisfies (5.10) for the H^1-norm in the setting of Example 3.1. To show this, we use the following result from [BPV00, Lemma 3.2]:

Lemma 5.6 *There exists a constant* C *not depending on* h_i, *such that*

$$\|u - I_i u\|_{L^2(\Omega)} \leq C h_i \|u\|_{H^1(\Omega)} \text{ for all } u \in H_0^1(\Omega).$$

Using similar techniques as in the proof of [Car99, Thm. 3.1], one can also show that

$$\|I_i u\|_{H^1(\Omega)} \leq C \|u\|_{H^1(\Omega)} \text{ for all } u \in H_0^1(\Omega).$$

These result are also valid for spaces where we have homogeneous Dirichlet conditions just on a part Γ_D of the complete boundary $\partial\Omega$, as long as it aligns with the meshes. That means that each edge of the triangulation is either contained in Γ_D or intersects Γ_D at most at the endpoint of the edge.

For the next lemma, we need the inverse estimate

$$\|v_i\|_{H^1(\Omega)} \leq h_i^{-1} \|v_i\|_{L^2(\Omega)} \quad \text{for all } v_i \in \mathcal{V}_i, \tag{5.14}$$

which is satisfied if the triangulation of Ω is quasi-uniform.

Lemma 5.7 *In the setting of Example 3.1, the quasi-interpolants* I_i, $i = 1, \ldots, r$ *satisfy* (5.10) *for* $\|\cdot\|_\# = \|\cdot\|_{H^1(\Omega)}$.

PROOF Let $i \le j < r$ and $v_{j+1} \in \mathcal{V}_{j+1}$. We set $I_j^i = I_i I_{i+1} \cdots I_j$. Using (5.14), (5.12), and the estimates above, we obtain

$$\begin{aligned}
\|I_j^i v_{j+1}\|_{H^1(\Omega)} &\le \|I_i(I_{i+1}(\cdots(I_j v_{j+1} - v_{j+1})))\|_{H^1(\Omega)} + \|(I_i(\cdots(I_{j-1} v_{j+1})))\|_{H^1(\Omega)} \\
&\le h_i^{-1} \|I_i(I_{i+1}(\cdots(I_j v_{j+1} - v_{j+1})))\|_{L^2(\Omega)} + \|I_{j-1}^i v_{j+1}\|_{H^1(\Omega)} \\
&\le h_i^{-1} \|I_j v_{j+1} - v_{j+1}\|_{L^2(\Omega)} + + \|I_{j-1}^i v_{j+1}\|_{H^1(\Omega)} \\
&\le C h_i^{-1} h_j \|v_{j+1}\|_{H^1(\Omega)} + \|I_{j-1}^i v_{j+1}\|_{H^1(\Omega)} \\
&\le C \sum_{k=i}^{j} h_i^{-1} h_k \|v_{j+1}\|_{H^1(\Omega)} = C \|v_{j+1}\|_{H^1(\Omega)} \sum_{k=i}^{j} h_i^{-1} h_k.
\end{aligned}$$

The sum in the last expression is bounded by means of the geometric series since $h_1 = 2^{j-1} h_j$ holds. This shows the assertion. □

5.3. Example 2

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with Lipschitz-boundary. We consider the problem

$$\begin{aligned}
\min_{u \in \mathcal{C}} J_2(u), \quad J_2 \colon H^1(\Omega) \cap L^\infty(\Omega) \to \mathbb{R}, \quad u \mapsto \int_\Omega j_2(x, u(x), \nabla u(x)) \, \mathrm{d}x, \\
j_2 \colon \Omega \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}, \; j_2(x, u, z) := \frac{1}{2}(\psi(u) z^T z + b(x) u^2)
\end{aligned} \tag{5.15}$$

with $\psi : \mathbb{R} \to \mathbb{R}_+$ twice continuously differentiable and $b \in L^\infty(\Omega)$ non-negative. We further demand that a lower bound $c_\psi > 0$ exists such that

$$\psi(u) \ge c_\psi \quad \text{for all } u \in \mathbb{R}.$$

As in the previous example we assume $\mathcal{C}$ to be a nonempty, closed and convex set that satisfies (5.3).

Remark 5.4 The problem is well-defined and possesses a solution if we use $H^1(\Omega)$ as domain of definition of J_2. This follows from Theorem A.7 and the coercivity of J_2, which can be shown by a straightforward calculation. However, in this setting we cannot show that J_2 satisfies the necessary differentiability assumptions. Hence, we consider instead $H^1(\Omega) \cap L^\infty(\Omega)$ as preimage space. This is justified by the fact that in typical cases, which depend on the feasible set $\mathcal{C}$, the solution has higher regularity and thus is an element of $L^\infty(\Omega)$. See for instance Theorem A.8 for the case with Dirichlet boundary conditions or pointwise bounds.

We first show that J_2 satisfies the differentiability assumptions (H1').

Lemma 5.8 *The functional J_2 is twice Gâteaux differentiable on $H^1(\Omega) \cap L^\infty(\Omega)$. Furthermore, the first derivative is continuous and the operator $u \mapsto J_2''(u)[d, d]$ is continuous for every fixed direction $d \in H^1(\Omega) \cap L^\infty(\Omega)$.*

PROOF Since ψ is twice continuously differentiable, there exists a constant $C(R) \geq 1$ for each $R \geq 0$ such that there hold

$$|\psi(u)| \leq C(R),\ |\psi'(u)| \leq C(R) \text{ and } |\psi''(u)| \leq C(R) \quad \text{for all } u \in \mathbb{R} \text{ with } |u| \leq R.$$

Without loss of generality we assume $C(R) \geq \|b\|_{L^\infty} \max\{R, R^2\}$. From the definition of j_2 we obtain the following estimates for $|u| \leq R$:

$$\begin{aligned}
|j_2(x,u,z)| &\leq C(R)\|z\|^2 + \|b\|_{L^\infty(\Omega)} R^2 \leq C(R)\tilde{V}^2, \\
\Big|\frac{\partial}{\partial u} j_2(x,u,z)\Big| &= \frac{1}{2}|\psi'(u) z^T z| + b(x)u \leq C(R)\|z\|^2 + \|b\|_{L^\infty(\Omega)} R \leq C(R)\tilde{V}^2, \\
\Big\|\frac{\partial}{\partial z} j_2(x,u,z)\Big\| &= \|\psi(u)z\| + \frac{1}{2}a(x)u^2 \leq C(R)\|z\| + \|b\|_{L^\infty(\Omega)} R^2 \leq C(R)\tilde{V},
\end{aligned}$$

where $\tilde{V} := (1 + \|z\|^2)^{1/2}$. Thus, the functional J_2 satisfies assumptions (A.3) of Theorem A.4 and it follows that J_2 is continuously differentiable. In the same way, one shows that the second-order partial derivatives of j_2 satisfy (A.6). Then the second-order differentiability of J_2 follows directly from Theorem A.5. □

In the previous chapters we have often considered the case $\mathcal{V} = H^1(\Omega)$. One important example is the equivalence of the multilevel stationarity measure and the dual-norm of the gradient, which greatly reduces the computational complexity. It is not evident that we obtain the same results for the space $\mathcal{V} = H^1(\Omega) \cap L^\infty(\Omega)$ or, more precisely, for the discrete finite element spaces $(\mathcal{V}_i, \|\cdot\|_{H^1 \cap L^\infty})$. However, in the discrete setting we can still work in $(\mathcal{V}_i, \|\cdot\|_{H^1})$ because the spaces are finite dimensional and the norms therefore equivalent. We would expect though that J_2' and hence the stationarity measure is not level-independently continuous anymore. The next lemma shows that this is indeed true, but the level-dependence is rather weak and of the same order as the one we have observed in the constrained case (cf. Section 4.3.1).

Lemma 5.9 *Let $(\mathcal{V}_h, \|\cdot\|_{L^\infty \cap H^1})$ be the space of continuous, piecewise linear functions defined on a quasi-uniform triangulation of $\Omega \subset \mathbb{R}^2$ with maximum diameter h. Then for every $g \in X^*$, $X := L^\infty(\Omega) \cap H^1(\Omega)$, there exists an element $g_h \in \mathcal{V}_h$ such that $(g_h, v_h) = \langle g, v_h \rangle$ for all $v_h \in \mathcal{V}_h$ and*

$$\frac{1}{C}\|\iota_{L^2}(g_h)\|_{X^*} \leq \|\iota_{L^2}(g_h)\|_{\mathcal{V}_h^*} := \sup_{v_h \in \mathcal{V}_h} \frac{(g_h, v_h)}{\|v_h\|_X} \leq \|\iota_{L^2}(g_h)\|_{X^*}$$

holds with a constant C that is independent of h and g. Furthermore, we have

$$\|g_h\|_{\mathcal{V}_h^*} \leq \|g_h\|_{(\mathcal{V}_h, \|\cdot\|_{H^1(\Omega)})^*} := \sup_{v_h \in \mathcal{V}_h} \frac{(g_h, v_h)}{\|v_h\|_{H^1(\Omega)}} \leq (1 + C|\log(h)|^{1/2})\|g_h\|_{\mathcal{V}_h^*}$$

for all $g_h \in \mathcal{V}_h$.

PROOF The existence of g_h with the asserted properties follows from the Riesz representation theorem. Using the stability of the L^2-projection with respect to $\|\cdot\|_{H^1(\Omega)}$ and $\|\cdot\|_{L^\infty}$, the equivalence of the norms on $\mathcal{V}_h$ follows as in Lemma 3.2.

The left-hand side of the second assertion follows directly from $\|v_h\|_{H^1(\Omega)} \leq \|v_h\|_{H^1(\Omega)\cap L^\infty(\Omega)}$. In order to verify the right-hand side, we use the well-known inequality $\|v_h\|_{L^\infty(\Omega)} \leq C|\log h|^{1/2}\|v_h\|_{H^1(\Omega)}$ (cf. [BS08, Lemma 4.9.2]) and obtain

$$\sup_{v_h\in\mathcal{V}_h} \frac{(g_h, v_h)}{\|v_h\|_X} \geq \sup_{v_h\in\mathcal{V}_h} \frac{(g_h, v_h)}{(1+C|\log h|^{1/2})\|v_h\|_{H^1(\Omega)}} = (1+C|\log h|^{1/2})^{-1}\|g_h\|_{(\mathcal{V}_h,\|\cdot\|_{H^1(\Omega)})^*},$$

which shows the assertion. □

The following corollary follows directly from the previous lemma:

Corollary 5.1 *Let $F\colon X \to X^*$, $X := H^1(\Omega)\cap L^\infty(\Omega)$, be uniformly continuous on a set $\mathcal{S} \subset \mathcal{V}_h$, i.e., for all $\varepsilon > 0$ there exists $\delta(\varepsilon) > 0$ such that*

$$\|F(v_h) - F(u_h)\|_{X^*} \leq \varepsilon \quad \text{for all } u_h, v_h \in \mathcal{S} \text{ with } \|u_h - v_h\|_X \leq \delta(\varepsilon),$$

with the space $\mathcal{V}_h$ as in the previous lemma. Then $F_h\colon (\mathcal{V}_h, H^1(\Omega)) \to (\mathcal{V}_h, H^1(\Omega))^*$ with $F_h(u_h) = F(u_h)$ for all $u_h \in \mathcal{V}_h$ is uniformly continuous on $\mathcal{S}$, more precisely for all $\varepsilon > 0$ it holds

$$\|F_h(v_h) - F_h(u_h)\|_{(\mathcal{V}_h,H^1(\Omega))^*} \leq \varepsilon \quad \text{for all } u_h, v_h \in \mathcal{S} \text{ with } \|u_h - v_h\|_{H^1(\Omega)} \leq \delta_h(\varepsilon)$$

where $\delta_h(\varepsilon) \leq \delta((1+C|\log(h)|^{1/2})\varepsilon)$.

In what follows, we assume that the function ψ and its derivatives are bounded on the feasible set, i.e., there exists a constant C_ψ such that

$$\|\psi^{(k)}(u)\|_{L^\infty(\Omega)} \leq C_\psi \quad \text{for all } u \in \mathcal{C} \text{ and } k = 0, 1, 2.$$

A simple calculation shows that the Gâteaux derivative of J_2 in direction $d \in X := H^1(\Omega)\cap L^\infty(\Omega)$ is given by

$$J_2'(u)[d] = \frac{1}{2}\int_\Omega \Big(\psi'(u)\nabla u^T\nabla u\cdot d + 2\psi(u)\nabla u^T\nabla d + b(x)ud\Big)\,\mathrm{d}x,$$

and its second derivative in directions $d_1, d_2 \in X$ by

$$J_2''(u)[d_1,d_2] = \frac{1}{2}\int_\Omega \Big(\psi''(u)d_1d_2\nabla u^T\nabla u + 2\big(\psi'(u)(d_1\nabla u^T\nabla d_2 + d_2\nabla u^T\nabla d_1) + \psi(u)\nabla {d_2}^T\nabla d_1\big)\Big)\,\mathrm{d}x.$$

Using Hölder's inequality, we get the estimate

$$\begin{aligned} J_2''(u)[d_1,d_2] &\leq \frac{1}{2}\|\psi''(u)\|_{L^\infty}\|d_1\|_{L^\infty}\|d_2\|_{L^\infty}\|\nabla u\|_{L^2}^2 \\ &\quad + \|\psi'(u)\|_{L^\infty}\|\nabla u\|_{L^2}(\|d_1\|_{L^\infty}\|\nabla d_2\|_{L^2} + \|d_2\|_{L^\infty}\|\nabla d_1\|_{L^2}) \\ &\quad + \|\psi(u)\|_{L^\infty}\|\nabla d_1\|_{L^2}\|\nabla d_2\|_{L^2} \qquad (5.16) \\ &\leq \frac{1}{2}(\|\psi''(u)\|_{L^\infty}\|\nabla u\|_{L^2}^2 + 2\|\psi'(u)\|_{L^\infty}\|\nabla u\|_{L^2} + 2\|\psi(u)\|_{L^\infty})\|d_1\|_X\|d_2\|_X \\ &\leq 2C_\psi(\|\nabla u\|_{L^2}^2 + 1)\|d_1\|_X\|d_2\|_X. \end{aligned}$$

On the sublevel sets

$$L^-_{\hat{u}}(J_2) := \{u \in H^1(\Omega) \cap L^\infty(\Omega) \,|\, J_2(u) \le J_2(\hat{u})\},$$

we have

$$\frac{1}{2} c_\psi \|\nabla u\|^2_{L^2(\Omega)} \le J_2(u) \le J_2(\hat{u}) \tag{5.17}$$

and hence the boundedness of the elements in $L^-_{\hat{u}}(J_2)$ with respect to the $H^1(\Omega)$-seminorm. By (5.4) it follows that also their $H^1(\Omega)$-norm is bounded. Thus, if we use a trust-region norm which satisfies $\|\cdot\|_i \ge \|\cdot\|_{H^1(\Omega)}$, $i = 1, \dots, r$, and suitable restriction operators, we can apply Lemma 5.4 with $\|\cdot\|_\# = \|\cdot\|_{H^1(\Omega)}$. This yields the boundedness of all iterates $x_i + v_{i,k} + t s_{i,k}$, $t \in [0,1]$, $i = 1, \dots, r$, $k = 1, 2, \dots$, in terms of $\|\cdot\|_{H^1(\Omega)}$. Together with (5.16) we finally conclude that there exists a level-independent constant C_H such that

$$J_2''(x_i + v_{i,k})[d_1, d_2] \le C_H \|d_1\|_X \|d_2\|_X. \tag{5.18}$$

Since $d = 2$ and the triangulation is quasi-uniform, there exists a level-independent constant C such that the inverse inequality

$$\|v_i\|_{L^\infty(\Omega)} \le C h_i^{-1} \|v_i\|_{L^2(\Omega)} \text{ for all } v_i \in \mathcal{V}_i, i = 1, \dots, r$$

holds (cf. for instance [Cia78, Thm. 3.2.6]). Hence, the following estimate is true:

$$\|v_i\|_X = \|\nabla v_i\|_{L^2(\Omega)} + \|v_i\|_{L^\infty(\Omega)} \le C h_i^{-1} \|v_i\|_{L^2(\Omega)}.$$

Estimating the norm in (5.18) by the last inequality yields

$$J_2''(x_i + v_{i,k})[d_1, d_2] \le C_H h_i^{-2} \|d_1\|_{L^2(\Omega)} \|d_2\|_{L^2(\Omega)} = C_H \lambda_i^{\max} \|d_1\|_{L^2(\Omega)} \|d_2\|_{L^2(\Omega)}$$

for all directions $d_1, d_2 \in \mathcal{V}_i$. Thus, (H4') is satisfied.

From (5.18) it also follows that (H2') and (H3') is satisfied if we choose $\|\cdot\|_X$ as trust-region norm on every level.

Finally, we verify that J_2' is uniformly continuous on $L^-_{\hat{v}}(J_2)$ by showing that it is Lipschitz continuous. By definition of the dual norm we have

$$\begin{aligned}
\|J_2'(u) - J_2'(v)\|_{X^*} &= \frac{1}{2} \sup_{\|d\|_X = 1} \Big[\int_\Omega (\psi'(u)\|\nabla u\|^2 - \psi'(v)\|\nabla v\|^2) d \,\mathrm{d}x \\
&\qquad + \int_\Omega \big(2(\psi(u)\nabla u - \psi(v)\nabla v)^T \nabla d + b(x)(u - v)d\big) \,\mathrm{d}x \Big] \\
&\le \frac{1}{2} \int_\Omega |\psi'(u)\|\nabla u\|^2 - \psi'(v)\|\nabla v\|^2| \,\mathrm{d}x \\
&\qquad + \|\psi(u)\nabla u - \psi(v)\nabla v\|_{L^2} + \frac{1}{2}\|b\|_{L^2}\|u - v\|_{L^2}.
\end{aligned} \tag{5.19}$$

In order to make further estimates, we first reformulate the integral term

$$\begin{aligned}
\int_\Omega |\psi'(u)\|\nabla u\|^2 - \psi'(v)\|\nabla v\|^2| \,\mathrm{d}x \le \frac{1}{2} \Big[&\int_\Omega |(\psi'(u) - \psi'(v))(\|\nabla u\|^2 + \|\nabla v\|^2)| \,\mathrm{d}x \\
&+ \int_\Omega |(\psi'(u) + \psi'(v))(\|\nabla u\|^2 - \|\nabla v\|^2)| \,\mathrm{d}x \Big].
\end{aligned}$$

Using the fundamental theorem of calculus for almost each $x \in \Omega$ then yields

$$\begin{aligned}\int_\Omega |\psi'(u) - \psi'(v)|(\|\nabla u\|^2 + \|\nabla v\|^2)\,\mathrm{d}x &= \int_\Omega \Big| \int_0^1 \psi''(u + t(v-u))(v-u)(\|\nabla u\|^2 + \|\nabla v\|^2)\,\mathrm{d}t\Big|\,\mathrm{d}x \\ &\le C_\psi \int_\Omega |v-u|(\|\nabla u\|^2 + \|\nabla v\|^2)\,\mathrm{d}x \\ &\le C_\psi \|v-u\|_{L^\infty}(\|\nabla u\|_{L^2}^2 + \|\nabla v\|_{L^2}^2) \le \frac{4C_\psi}{c_\psi} J_2(\hat{u})\|v-u\|_X.\end{aligned}$$

We have used that $\mathcal{C}$ is a convex set and that (5.17) holds for all $u, v \in L_{\hat{u}}^-(J_2)$. Further, we have

$$\begin{aligned}\int_\Omega (\psi'(u) + \psi'(v))|\|\nabla u\|^2 - \|\nabla v\|^2|\,\mathrm{d}x &\le 2C_\psi \int_\Omega |(\nabla u - \nabla v)^T(\nabla u + \nabla v)|\,\mathrm{d}x \\ &\le 2C_\psi \|\nabla u + \nabla v\|_{L^2}\|\nabla u - \nabla v\|_{L^2} \\ &\le 4C_\psi \Big(\frac{2}{c_\psi} J_2(\hat{u})\Big)^{1/2}\|u-v\|_X.\end{aligned}$$

Similarly, we estimate the second term in (5.19):

$$\begin{aligned}\|\psi(u)\nabla u - \psi(v)\nabla v\|_{L^2} &\le \frac{1}{2}\Big[\|(\psi(u) - \psi(v))(\nabla u + \nabla v)\|_{L^2} + \|(\psi(u) + \psi(v))(\nabla u - \nabla v))\|_{L^2}\Big] \\ &\le \frac{1}{2}\Big(\frac{2}{c_\psi} J_2(\hat{v})\Big)^{1/2}\|\psi(u) - \psi(v)\|_{L^\infty} + C_\psi\|\nabla u - \nabla v\|_{L^2} \\ &\le \frac{1}{2} C_\psi \Big(\frac{2}{c_\psi} J_2(\hat{v})\Big)^{1/2}\|u-v\|_X + C_\psi\|u-v\|_X.\end{aligned}$$

Inserting all estimates in (5.19) shows the Lipschitz continuity of J_2' on $L_{\hat{u}}^-(J_2)$ and thus (H5).

5.4. Minimum surface problems

Let $\Omega \subset \mathbb{R}^2$ be a Lipschitz-continuous domain with boundary Γ. Furthermore let u_0 be a continuous function on Γ that describes the values of a surface on the boundary. The solution of the problem $\min_{u\in\mathcal{C}} J_3(u)$ with $\mathcal{C} := \{u \in H^1(\Omega) |\, (u - u_0) \in H_0^1(\Omega)\}$ and

$$J_3(u) := \int_\Omega \sqrt{1 + \nabla u^T \nabla u}\,\mathrm{d}x$$

describes the minimum surface. Whether a solution of this problem exists depends on the domain Ω and the set $\mathcal{C}$. It is well known that the problem has a solution if Ω is convex and u_0 is a $C^2(\Omega)$-function (cf., e.g., [Giu03, Theorem 1.6]). As a variant, we also consider different feasible sets $\mathcal{C}$ where besides the Dirichlet boundary conditions additional constraints on the surface are demanded, as for example that it has to lie above an obstacle. In this case, however, the solvability of the problem is not so easy to analyse. If we replace $H^1(\Omega)$ by a suitable finite element space $\mathcal{V}_h$ with mesh size h, it is easy to see that the problem is always solvable. However, the solutions will in general not converge for $h \to 0$.

To show that J_3 satisfies (H1'), we define the C^2-function $j_3(x,u,z) := \sqrt{1+z^Tz}$ and estimate the partial derivatives of j_3:

$$\begin{aligned} |j_3(x,u,z)| &\leq 1 + \|z\|, \\ \Big\|\frac{\partial}{\partial z} j_3(x,u,z)\Big\| &= \frac{\|z\|}{\sqrt{1+z^Tz}} \leq 1, \\ \Big\|\frac{\partial^2}{\partial z \partial z} j_3(x,u,z)\Big\| &= \frac{1}{\sqrt{1+z^Tz}} - \frac{\|zz^T\|}{(1+z^Tz)^{3/2}} \leq 2. \end{aligned}$$

Hence, assumption (A.2) of Theorem A.4 and assumption (A.5) of Theorem A.5 are satisfied, which shows (H1') with $\mathcal{V} = H^1(\Omega)$.

The second-order directional derivative is given by

$$J_3''(u)[d,d] = \int_\Omega \frac{\nabla d^T \nabla d (1 + \nabla u^T \nabla u) - (\nabla u^T \nabla d)^2}{(1+\nabla u^T \nabla u)^{3/2}} \,\mathrm{d}x.$$

We can easily derive an upper bound for $J_3''(u)[d,d]$ in terms of $\|d\|_{H^1}$:

$$\begin{aligned} |J_3''(u)[d,d]| &\leq \int_\Omega \Big(\|\nabla d\|^2 + \frac{(\nabla u^T \nabla d)^2}{(1+\nabla u^T \nabla u)^{3/2}} \Big) \,\mathrm{d}x \leq \int_\Omega \Big(\|\nabla d\|^2 + \frac{\|\nabla u\|^2 \|\nabla d\|^2}{1+\nabla u^T \nabla u} \Big) \,\mathrm{d}x \\ &\leq 2\|\nabla d\|^2_{L_2(\Omega)} \leq 2\|d\|^2_{H^1(\Omega)} \end{aligned}$$

This verifies assumptions (H2'), (H3') and (H4') if $\|\cdot\|_{H^1(\Omega)}$ or the H^1-semi-norm is chosen as trust-region norm. Moreover, since the operator $J_3''(u)$ is bounded by $L = 2$ in $\mathcal{L}(H^1(\Omega), (H^1(\Omega))^*)$ for all $u \in H^1(\Omega)$, it follows from Lemma A.4 that J_3' is Lipschitz continuous on $H^1(\Omega)$ and hence satisfies (H5).

5.5. Signorini Problem

The *Signorini problem* is a simple contact problem from the theory of linear elasticity. A deformation of an elastic body is searched that is subjected to body forces and surface tractions and which has frictionless contact to a rigid obstacle on some part of his surface. The contact area is not known in advance but is part of the solution. Instead of the whole nonlinear elasticity model, a linearization is used and hence the results are only valid for small deformations of the body. We shall give just a very short description of the problem, a more comprehensive introduction can be found for instance in [KO88, Ch. 2 & 6].

Assume $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, is a domain with Lipschitz-boundary. We will think of $\overline{\Omega}$ as the part of space occupied by a body in a natural state, i.e., unstressed state, before it is deformed. We assume that the boundary $\partial\Omega$ consists of three parts $\Gamma_D, \Gamma_N, \Gamma_C$ with $\Gamma_D \cap \Gamma_C = \emptyset$. We have Dirichlet conditions on Γ_D and traction forces $t \in \big(L^\infty(\Gamma_N)\big)^d$ act on Γ_N. The area of the boundary where contact to the obstacle is possible is denoted by Γ_C. For simplicity, we demand

$\Gamma_D \neq \emptyset$[2]. We are looking for a displacement $v^* \in \mathcal{V} := \{v \in H^1(\Omega)^d \,|\, v = 0 \text{ on } \Gamma_D\}$ that solves the problem

$$\begin{aligned} &\min_{v \in \mathcal{V}} \int_\Omega \left[\mu \varepsilon(v) : \varepsilon(v) + \frac{\lambda}{2} (\operatorname{div} v)^2 - f \cdot v \right] \mathrm{d}x - \int_{\Gamma_N} t \cdot v \, \mathrm{d}S(x) \\ &\text{s.t. } v^T n \leq g \text{ on } \Gamma_C. \end{aligned} \tag{5.20}$$

Here, $f \in (L^\infty(\Omega))^d$ denotes the body forces, $\varepsilon(v) := \frac{1}{2}(\nabla v + \nabla v^T)$ the linearized strain tensor and λ and μ are material parameters (Lamé's parameters). $n(x) \in \mathbb{R}^d$ denotes the normal vector at $x \in \Gamma_C$ and $g(x) \geq 0$ the Euclidean distance (gap) from x to the rigid obstacle in direction $n(x)$.

The set $\mathcal{C} := \{v \in \mathcal{V} \,|\, v^T n \leq g \text{ a.e. on } \Gamma_C\}$ of feasible displacements is closed and convex.

As it is shown for instance in [KO88, Theorem 6.1], the problem admits a unique solution if $\mathcal{C} \neq \emptyset$.

The objective function of the Signorini problem (5.20) is quadratic and hence it is easily seen that it satisfies all assumptions (H1')–(H4') and (H5) for the trust-region norms $\|\cdot\|_i = \|\cdot\|_{H^1(\Omega)^d}$.

5.5.1. Discretization

We will now shortly introduce a finite element discretization of Signorini's problem based on [KK01] that leads to a box constrained minimization problem.

We assume that $\overline{\Omega}$ is polygonal. Let $\mathcal{T}_1$ be a quasi-uniform triangulation (tetrahedra in $\mathbb{R}^3$) of $\overline{\Omega}$ with minimum diameter h_1 and let $\mathcal{N}_1$ be the corresponding set of free nodes $x_1^1, \ldots, x_1^{n_1}$, i.e., all the nodes that are not contained in Γ_D. We assume that Γ_D aligns with the triangulations, i.e., an edge is either contained in Γ_D or intersects it at most at the endpoint of the edge.

Let $\phi_1^i : \Omega \to \mathbb{R}$, $i = 1, \ldots, n_1$, be the piecewise linear nodal basis functions that satisfy $\phi_1^i(x_1^k) = \delta_{ik}$ for all nodes $x_1^k \in \mathcal{N}_1$. The finite element space is now defined by

$$\mathcal{V}_1 := \left\{ v_1 = \sum_{i=1}^{n_1} \tilde{v}_{1,i}^c \phi_1^i \,|\, \tilde{v}_{1,i}^c \in \mathbb{R}^d,\ i = 1, \ldots, n_1 \right\} \subset \mathcal{V},$$

where $\tilde{v}_1^c \in \mathbb{R}^{n_1 \cdot d}$ is the corresponding coefficient vector of v_1. A standard approximation of the contact condition is to demand it only for nodes on the contact boundary, i.e.,

$$v_1(x_1^k)^T n(x_1^k) \leq g(x_1^k) \Leftrightarrow (\tilde{v}_{1,k}^c)^T n(x_1^k) \leq g(x_1^k) \text{ for all } x_1^k \in \mathcal{N}_1 \cap \Gamma_C. \tag{5.21}$$

In [HL77] it was shown that the solution of this finite element approximation converges to the solution of (5.20) for mesh sizes $h \to 0$.

In general, (5.21) does not lead to standard box conditions on the coefficient vectors. In order to formulate it as a simple bound constraint, we use a special local orthogonal basis of $\mathbb{R}^d$ for

[2] If no Dirichlet boundary is given, additional conditions are needed to assure unique solvability of the problem, cf. [KO88].

every grid point in $x_1^k \in \mathcal{N}_1$. This basis is represented by a matrix $Q(x_1^k) \in \mathbb{R}^{d\times d}$ whose first column is equal to $n(x_1^k)$ for $x_1^k \in \Gamma_C \cap \mathcal{N}_1$. One possibility to construct $Q(x_1^k)$ is by rotating the Cartesian basis such that the first unit vector ended up on $n(x_1^k)$ (Givens rotation). On grid points that are not contained in Γ_C, an arbitrary orthogonal basis, e.g., the standard Cartesian basis, $Q(x_1^k) = I$, can be chosen. With this, we obtain a different representation of the functions in $\mathcal{V}_1$:

$$v_1 = \sum_{i=1}^{n_1} Q(x_1^i)\tilde{v}_{1,i}\phi_1^i, \quad \tilde{v}_{1,i} \in \mathbb{R}^d.$$

The contact condition then becomes

$$(Q(x_k)\tilde{v}_{1,k})^T n(x_k) = (\tilde{v}_{1,k})^1 \leq g(x_k) \text{ for } x_k \in N_1 \cap \Gamma_C,$$

which is a simple upper bound on the first entry of each part of the coefficient vector. We define

$$Q_1 := \begin{pmatrix} Q(x_1^1) & & & \\ & Q(x_1^2) & & \\ & & \ddots & \\ & & & Q(x_1^{n_1}) \end{pmatrix}$$

and $\Phi_1 := (\phi_1^1 e^1, \phi_1^1 e^2, \ldots, \phi_1^1 e^d, \ldots, \phi_1^{n_1} e^1, \ldots \phi_1^{n_1} e^d)$, where $e^j \in \mathbb{R}^d$, $j = 1, \ldots, d$, denotes the j-th unit vector. Then each element of $\mathcal{V}_1$ can be written as $v_1 = \Phi_1 Q_1 \tilde{v}_1$. The finite dimensional problem in terms of the new coefficient vector is

$$\begin{aligned} \min_{\tilde{v}_1 \in \mathbb{R}^{d\cdot n_1}} \ & \frac{1}{2}\tilde{v}_1^T Q_1^T C_1 Q_1 \tilde{v}_1 - f_1^T Q_1 \tilde{v}_1 - g_1^T Q_1 \tilde{v}_1 \\ \text{s.t. } & \tilde{v}_{1,k}^1 \leq g(x_1^k) \text{ for } x_1^k \in \mathcal{N}_1 \cap \Gamma_C \end{aligned}$$

where

$$(C_1)^{ij} = \int_\Omega \left(2\mu\varepsilon(\Phi_1^i) : \varepsilon(\Phi_1^j) + \lambda \mathrm{div}(\Phi_1^i)\mathrm{div}(\Phi_1^j)\right) \mathrm{d}x,$$

$f_1^i = (f, \Phi_1^i)_{L^2(\Omega)}$ and $g_1^i = \int_{\Gamma_N} g\Phi_1^i \, \mathrm{d}S(x)$. Since Q_1 is an orthogonal matrix, the condition of the problem is not influenced by the transformation.

For the multilevel algorithm, we construct the spaces $\mathcal{V}_2 \subset \mathcal{V}_3 \subset \cdots \subset \mathcal{V}_r$ in the same way where the underlying triangulations of $\overline{\Omega}$ are obtained by uniform refinement.

5.6. Nonlinear elasticity

Often the linearized elasticity model used in the previous section are not accurate enough, e.g., for large displacements of the body. In this case, one has to work with nonlinear models. A special class of nonlinear materials are *hyperelastic* materials, for which a stored energy density function $\hat{W} : \overline{\Omega} \times \mathbb{R}^{d\times d} \to \mathbb{R}$ exists. Roughly speaking, $\hat{W}$ assigns each point of the reference configuration and each deformation gradient the strain energy in this point. A typical example of a hyperelastic material is rubber.

Given body forces f and forces g that act on the Neumann parts of the boundary, the total potential energy of the body is given by

$$J_5(v) := \int_\Omega \left(\hat{W}(x, F) - f \cdot v\right) \mathrm{d}x - \int_{\Gamma_N} g \cdot v \, \mathrm{d}S(x),$$

where $F(x) := I + \nabla v(x)$ denotes the deformation gradient. Typically, one postulates that a rotation of the whole system does not change its outcome. This axiom is called *frame-indifference*. We further assume that the material is homogeneous, i.e., $\hat{W}$ depends only on the deformation gradient and not on x. In this case there exists a function $W \colon \mathbb{M}^d_+ \to \mathbb{R}$, $\mathbb{M}^d_+$ being the set of symmetric and positive definite $d \times d$ matrices, such that $W(C) = \hat{W}(x, F)$ where $C = F^T F$ denotes the *right Cauchy-Green strain tensor* [Cia88, Thm. 4.2-1].

Furthermore, it is necessary to constrain the space of possible displacements such that physical not possible deformations like self penetration can not happen. A mathematically suitable constraint is $\det F > 0$ almost everywhere on Ω.

There are many different models for hyperelastic materials and it would go beyond the scope of this thesis to discuss them in detail. Hence, we will consider only the special class of *Compressible Mooney-Rivlin* materials, which was suggested in [CG82] (c.f. also [Cia88, Chapter 4]). The stored energy function is given by

$$\hat{W}(F) = a\|F\|^2 + b\|\operatorname{cof} F\|^2 + \gamma(\det F) + e$$

or respectively in terms of the right Cauchy-Green strain tensor by

$$W(C) = a \operatorname{tr} C + b \operatorname{tr}(\operatorname{cof} C) + \gamma(\sqrt{\det C}) + e$$

with parameters $a, b > 0$, $\gamma(\delta) = c\delta^2 - d\log(\delta)$, $c, d > 0$ and $e \in \mathbb{R}$. Here, cof denotes the cofactor matrix. A common demand is that for small deformations the hyperelastic material reassembles the properties of the linear model. This restricts the choice of the parameters to

$$a = \frac{\mu}{2} + c - \frac{\lambda}{4}, \quad b = \frac{\lambda}{4} - c, \quad c < \frac{\lambda}{4}, \quad d = \frac{\lambda}{2} + \mu, \quad e = -(3a + 3b + c)$$

where $\lambda \in \mathbb{R}$ and $\mu > 0$ are the Lamé constants.

Although one can show that this stored energy function is not convex (this holds true for any reasonable non-linear material, cf. [Cia88, Thm. 4.8-1]), this material has the advantage that its stored energy function is polyconvex and one can show the existence of a solution, i.e., there exists at least one $v^* \in H^1(\Omega)^d$ such that $J_5(v^*) = \inf_{v \in H^1(\Omega)^d} J_5(v)$. A detailed discussion of this theory, which goes back to John Ball, can be found in [Cia88, Chapter 7]. Furthermore, the log-term serves also as an implicit barrier for the constraint $\det F > 0$.

It is not possible to show the necessary differentiability properties with the theory that we have used for the last examples. The problem is that an actual material has the property that an infinite amount of energy is required in order to annihilate volumes. Mathematically this can be expressed by the assumption that $\hat{W} \to \infty$ for $\det F \to 0^+$. Hence, the growth conditions which we have used in the previous cases do not hold.

To our knowledge, there is no satisfactory theory about the differentiability of the function J_5. Nonetheless, we can use our method since, in the discrete setting, the differentiability is

ensured. In the worst case, we will observe level-dependent factors when we increase the mesh size.

6. Numerical results

In this chapter, we apply Algorithm 2.1 to various 2D and 3D test problems which are mostly of the type discussed in the previous chapter. The algorithm allows a lot of freedom for the concrete choices of the parameters and the sub-algorithms used. Therefore, we will first introduce two different concrete implementation variants. Then we will describe the test problems and analyze the numerical performance of the algorithm. We do not focus here on the absolute runtime of the algorithm but instead on its behaviour when the number of levels grow. Furthermore, we do not compare the algorithm with other general purpose optimization methods. This comparison would not be completely fair since we consider a special class of optimization problems that is well suited, and we would therefore expect that our algorithm clearly outperforms these codes. This presumption is confirmed by the results in [GMS$^+$10] where a multilevel optimization code was tested against a standard Newton trust-region algorithm. However, it would be interesting to compare the performance against Multigrid-Newton methods for optimization problems. We assume that this would be a much closer race.

6.1. Two variants of Algorithm 2.1

Standard multigrid algorithms typically use a fixed iteration cycle, i.e., a rule, only depending on the iteration number, which determines when the algorithm smooths and when it changes the level. Most commonly used are V- and W-cycles (cf. Figure 6.1).

These fixed cycles are in general not possible in our algorithm since we are only allowed to go on a coarser grid when the smoothing property

$$\chi_j(0) \geq \kappa_\chi \chi_i(v_{i,k}), \tag{6.1}$$

is satisfied. Furthermore, it follows from the theory in Chapter 3 and 4 that we cannot expect an adequate descent of a smoothing step if the iterate is already smooth. Nonetheless, a strategy similar to a V-cycle can be used and works quite well for a large class of examples. The following algorithm shows the concrete implementation of Step 1 and Step 5 in Algorithm 2.1 for a V-cycle

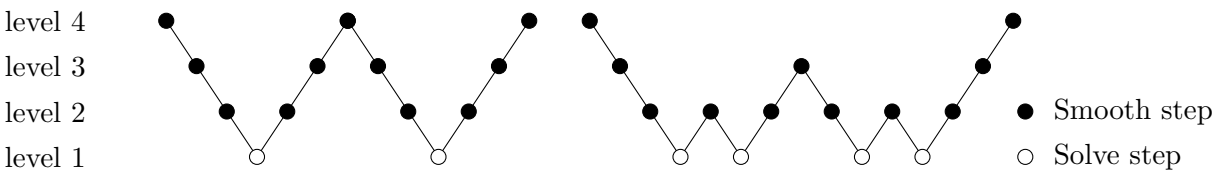

Figure 6.1.: Two V-cycles (left) and one W-cycle with pre- and postsmoothing

strategy with presmoothing. As in the proof of Lemma 2.10, we denote by $\theta(k)$ the number of successful iterations until the k-th iteration of the algorithm.

Algorithm 6.1 (TRMLConv(V-cycle))
Step 1: Model choice
If $i > 1$, $\theta(k) > 0$, (6.1) and

$$\chi_{i-1}(0) \geq \varepsilon_{i-1}^{\chi}$$

are satisfied, go to Step 2 (Multilevel step). Otherwise, go to Step 3.

Step 5: Termination
Return if one of the termination criteria in Step 5 of Alg. 2.1 is satisfied or if $i < r$ and one successful multilevel step was made. Otherwise, set $k \leftarrow k + 1$ and go to Step 1.

A different strategy that is better suited for our theory is to choose whether to make a smoothing step or enter a lower level depending on the smoothness of our current iterate. In comparison to the V-cycle version, we do not terminate automatically after a successful multilevel step on the lower levels but instead when the residuum is reduced suitably or a maximum number of successful iterations were made. The following algorithm shows how Step 1 and Step 5 are implemented.

Algorithm 6.2 (TRMLConv(Free))
Besides the parameters of Algorithm 2.1, this variant introduces two additional constants $0 < \kappa_{red} < 1$ and $k_{\max} \in \mathbb{N}$.

Step 1: Model choice
If (6.1), $i > 1$ and

$$\chi_{i-1}(0) \geq \varepsilon_{i-1}^{\chi}$$

are satisfied, go to Step 2 (Multilevel step). Otherwise, go to Step 3.

Step 5: Termination
Return if one of the termination criteria in Step 5 of Alg. 2.1 is satisfied. If $i < r$ and $\chi_i(v_{i,k+1}) \leq \kappa_{red}\chi_i(v_{i,0})$ or the number of successful iterations satisfies $\theta(k) \geq k_{\max}$, return with $v_{i,k+1}$. Otherwise set $k \leftarrow k + 1$ and go to Step 1.

For simple problems, the V-cycle algorithm performs often slightly better than the free form version. In these cases, the level pattern of the free form algorithm is similar to the V-cycle algorithm but with postsmoothing instead of presmoothing. One disadvantage of the free form version is that the smoothing parameter κ_χ must be chosen more carefully. If it is too low, convergence can slow down since virtually no smoothing steps are made. It could be interesting for future research to determine a good smoothing parameter automatically. In our tests, we use $\kappa_\chi = 0.7$ for the V-cycle and $\kappa_\chi = 0.8$ for the free form algorithm. The free form variant often leads to faster convergence for more complex problems, and one problem the V-cycle algorithm was not capable to solve within a reasonable number of function evaluations. In this cases, the free form version uses the lower levels more extensive and behaves more like a W-cycle algorithm.

6.2. Details of the implementation

We will now discuss some of the parameter and algorithmic selections we made. It would go beyond the scope of this thesis to numerically justify each choice in great detail. However, we have tried to identify the important parameters whose choices have major influence on the performance of the algorithm. This will later be illustrated on selected examples. These are chosen such that the observed effects are also representative for the majority – but not necessarily for all – of the other examples too.

6.2.1. Discretization

The feasible sets of the problems in this chapter are subsets of the infinite dimensional space $H^1(\Omega)$ where Ω is a polygonal domain. To calculate approximate solutions to this problem, we triangulate Ω and construct a hierarchy of finite element spaces with piecewise linear and continuous functions by uniform refinement of the grid as in Example 3.1. All Dirichlet boundary conditions are implicitly handled and do not occur as constraints in the discrete problems. This setting satisfies Assumptions 5.1 and the assumptions we made on the spaces in Chapter 3 and Chapter 4.

More details on the implementation of the prolongation and restriction operators, and the smoothing algorithm for the coefficient vectors can be found in Section 3.4.

As stationarity measure we use the multilevel stationarity measure χ_i^{ML} defined in (4.5), which is equivalent to the measure introduced in Theorem 3.4 if the problem is unconstrained.

We only consider constrained problems where we have pointwise bounds on the variables. This leads to box constraints on the coefficient vectors as described in Example 4.1. We use Algorithm 4.3 to create the lower-level boxes. We have obtained similar results using the construction of Lemma 4.10, though. To allow larger steps on the coarser levels we use the active-set strategy introduced in Section 4.3.2. This leads to a large performance increase in comparison to the standard version, which we will illustrate on some selected examples.

6.2.2. Hessian approximation

Standard multigrid algorithms for linear elliptic problems are known to converge with a linear rate. So – at best – we would also expect linear convergence for our nonlinear multigrid algorithm. This suggests that it is not necessary to always work with the exact Hessian in our quadratic models $q_{i,k}$. Instead, we use a heuristic strategy to update the Hessian that is similar to the strategy used in [GMTWM08]. For a Taylor step, we calculate a new Hessian if one of the following criteria is met:

1. The current level is the coarsest level, i.e., $i = 1$.
2. No previous Hessian approximation is available.
3. The previous iteration was a non-successful smoothing iteration.

4. The current Hessian approximation $H_{i,k}$ does not suitably describe the curvature in the direction of the last step. Suitably means here that for given $C_{HA} > 0$ and $\alpha \geq 1$ the inequality

$$\|\nabla h_i(v_{i,k}) - \nabla h_i(v_{i,k-1}) - H_{i,k}(v_{i,k} - v_{i,k-1})\| \leq C_{HA}\|v_{i,k} - v_{i,k-1}\|^{\alpha}$$

is violated.

Otherwise, we set $H_{i,k} = H_{i,k-1}$ for $k > 0$ or initialize $H_{i,0}$ with the approximation we have used the last time we visited this level. If we calculate a new Hessian on a level i, we also recalculate all Hessians on levels j with $j \prec i$ when they are needed the next time.

Since calculating the Hessian is in many cases by far the most expensive operation in our algorithm, this massively improves the performance. For our experiments, we have chosen $C_{HA} = 0.5$ and $\alpha = 3/2$.

6.2.3. Full multigrid

In our theory we have used the level hierarchy only to calculate correction steps. However, one can (and should) also use it to obtain a good initial iterate on the finest level. This strategy is often called *the full multigrid* or *nested iteration*. The idea is to successively solve the lower-level problems

$$\min_{v_i \in C_i} f_i(v_i)$$

for $i = 1, \ldots, r-1$ up to a certain precision and use the prolongated solution as initial value for the next finer level. The coarser problems can be solved cheaply and provide us with a good initial iterate. The feasible sets C_i are suitable approximations of the feasible set C_r here.

6.2.4. Trust-region radius update

To update the trust-region radius, we do not use the simple update rule (2.33) of Algorithm 2.1 but a more practical choice that was proposed in [CGT00, Ch. 17]:

$$\Delta_{i,k}^{+} := \begin{cases} \max\{\Delta_{i,k}, \gamma_1 \|s_{i,k}\|_i\} & \text{if } \rho_{i,k} \geq \eta_2, \\ \Delta_{i,k} & \text{if } \eta_1 \leq \rho_{i,k} < \eta_2, \\ \gamma_2 \min\{\Delta_{i,k}, \|s_{i,k}\|_i\} & \text{if } \rho_{i,k} < \eta_1, \end{cases}$$

with $\gamma_1 = 2$, $\gamma_2 = 0.5$, $\eta_1 = 0.1$ and $\eta_2 = 0.75$. This update rule does not suffer from the typical problem that the trust-region radius can become very large due to many very successful small steps and then needs a lot of unsuccessful iterations to be small enough to constrain the step length. The global convergence results of Chapter 2 remain valid with this trust-region update.

6.2.5. Smoother

We use Algorithm 4.2 with $m = 6$ and $\theta = 1.48$ as smoother to calculate the Taylor steps if $i > 1$. This algorithm seems to be most robust for both constrained and unconstrained problems.

In comparison to classic multigrid methods where the number of smoothing cycles is normally smaller than 3, the choice $m = 6$ seems rather large. This is justified by the fact that after a smoothing step we have to evaluate the objective function and – for the stationarity measure – its derivative at the new iterate. This is in general more expensive than a couple of smoothing cycles. Hence, we choose a larger number to minimize the number of function evaluations.

The choice of the relaxation parameter θ also has a large influence on the performance. In nearly all examples an overrelaxation increases the convergence speed, which we will illustrate on some examples.

6.2.6. Coarse grid solver

The degrees of freedom on the coarsest grid is typically very small and therefore we can use a more sophisticated algorithm to approximately solve the optimization problem, that uses the second-order information more extensively. We choose an affine scaling trust-region method [CL96] and use a standard Steihaug-Toint CG method (cf. [CGT00, Alg. 7.5.1]) for the calculation of the trial steps, which was very fast and reliable in our examples.

6.2.7. Termination criteria

For a typical user the multilevel stationarity measure is difficult to interpret. Hence, we use a more commonly used measure to decide when we terminate the iteration in Step 5: The projected gradient of the current step in the supremum norm

$$\chi_r^{\mathrm{ter}}(v_{r,k}) := \|v_{r,k} - \mathrm{Proj}_{C_r}(v_{r,k} - \nabla f_r(v_{r,k}))\|_\infty$$

where $\nabla f_r(v_{r,k})$ denotes the standard euclidean representation of $f_r'(v_{r,k})$. We terminate the algorithm if this measure is smaller than $\varepsilon_r^\chi = 10^{-8}$.

If not said otherwise, we use the same parameter set for all examples. Of course, this is not in every case the optimal choice, but it shows that the algorithm can be used for this kind of problems without tweaking the parameters.

6.2.8. Computational framework

To implement the algorithm, we use the platform independent language PYTHON with the linear algebra libraries NUMPY and SCIPY [JOP$^+$]. This language enjoys high acceptance in the scientific computing community – even for high performance computing – because it is possible to quickly implement algorithms in an interpreted language using high level structure like vectors and matrices with a good performance due to highly optimized libraries. To implement some time critical parts, like the smoothing algorithms, we use C++.

To calculate function values, gradients and Hessians of the more complex examples, we use the finite element toolbox FENICS [LMW$^+$11], which is programmed in C++ and provides an interface to PYTHON.

All tests were made on an Intel Xeon CPU with 2.93 GHz core speed. The code uses only one processor core.

6.3. Test problems

We have applied our algorithm to various test problems. We use some classical problems from the MINPACK-2 test problem collection [ACM91] and COPS[1] [DMM04] that are suited for our algorithm as well as some new examples. Since the total computation time is dominated by the time used for operations on the finest grid, the numbers in the result tables denote solely fine grid quantities as for example the number of function evaluations and multilevel steps.

We measured the time the algorithm needs for the optimization, including the nested iteration to obtain a good initial point, but without the time needed to create the level structure like the refined meshes and the prolongation operators. Unless otherwise stated, we use the V-cycle variant of the algorithm with the settings discussed previously and a $H^1(\Omega)$-trust-region.

6.3.1. Bound constrained quadratic problems

The objective functions of the first two problems, taken from [DMM04], are quadratic. In both cases the feasible set is given by pointwise bound constraints. The third problem is a 3D contact problem from linear elasticity.

Elasto-Plastic torsion problem

Let $\Omega \subset \mathbb{R}^2$ be a domain with Lipschitz-boundary. We consider an infinitely long cylindrical bar with cross section Ω that is made up of an isotropic elastic perfectly plastic material. Starting from a zero-stress initial state, an increasing torsion movement is applied. The constant $c > 0$

[1] Constrained Optimization Problem Set

Levels	dof	ML	Smooth	f	f'	time (sec.)
5	63^2	5	7	13	12	0.1
6	127^2	4	6	11	10	0.2
7	255^2	4	6	11	10	0.4
8	511^2	4	6	11	10	1.5
9	1023^2	5	6	12	11	6.0
5	63^2	11	12	24	23	0.3
6	127^2	12	13	26	25	0.4
7	255^2	12	13	26	25	0.9
8	511^2	10	11	22	21	2.5
9	1023^2	14	15	30	29	13.4

Table 6.1.: Results for Elasto-Plastic Torsion with $\theta = 1.48$ and active-set strategy (top) and with $\theta = 1$ and no active-set strategy (bottom).

characterises the torsion strength. The resulting stress potential v^* is the solution of the variational problem

$$\min_{v \in H_0^1(\Omega)} \frac{1}{2} \int_\Omega \|\nabla v\|^2 \,\mathrm{d}x - c \int_\Omega v \,\mathrm{d}x \quad \text{s.t. } |v(x)| \leq d(x, \partial\Omega) \text{ a.e. on } \Omega.$$

The corresponding stress field is then given by $\theta = \nabla v^*$. More details on this problem can be found for instance in [Glo84, Sec. II.3].

For our tests, we use the same problem parameters $\Omega = (0,1)^2$ and $c = 5$ as in [DMM04].

The first part of Table 6.1 shows the results of the optimization where the standard parameters were used. Each row corresponds to a full run with the given number of levels and degrees of freedom (*dof*). The entries in the columns labeled *ML* and *Smooth* show the number of multilevel steps and the number of smoothing iterations that were necessary on the finest grid before the algorithm terminates. Similar the entries in the columns f, f' and f'' show the number of function, gradient and Hessian evaluations on the finest grid. If the function is quadratic, we only need one Hessian evaluation and omit the entry f'' in the result table.

We can see that the algorithm needs roughly the same amount of work independent of the number of levels and the mesh-size of the discretization. The total computational time grows linearly with a factor of 4, which corresponds exactly to the increase in the number of unknowns. Hence, we have optimal complexity in this example.

To show the positive effect of the overrelaxation and the active-set strategy, we calculate the same example without these choices. The results in the second part of Table 6.1 show that the algorithm needs twice as much time in this case. Figure 6.2 also shows the positive effect of these choices.

Remark 6.1 We recall that our active-set method is similar to the truncated basis methods used in [Kor94] for monotone multigrid methods. There, a similar performance increase was numerically shown in comparison to the standard method, see also [GK09b].

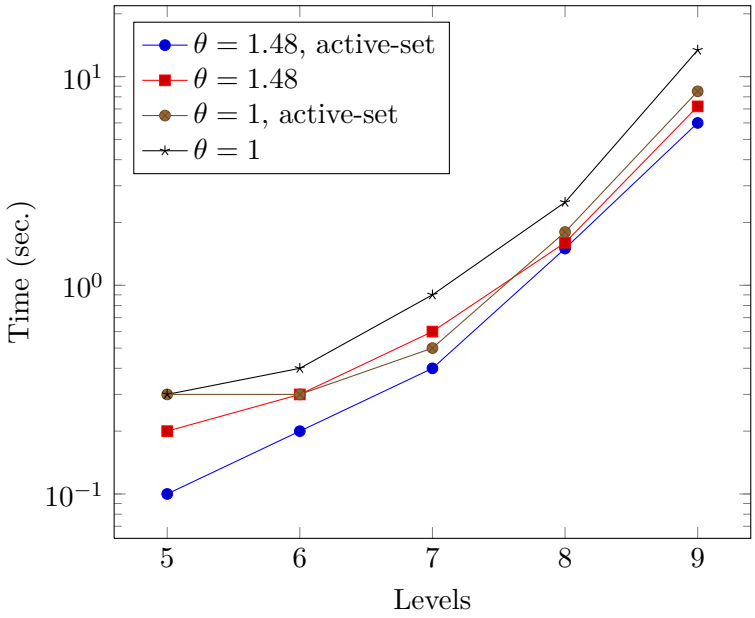

Figure 6.2.: Performance of the Elasto-Plastic-Torsion problem for different parameter choices

Journal-Bearing problem

The journal bearing problem simulates the pressure distribution between two circular cylinders of length L and radii R and $R + c$. The separation between the cylinders is εc, where $0 \leq \varepsilon < 1$ is the eccentricity. The pressure is the solution of the problem

$$\min_{v \in H_0^1(\Omega)} \frac{1}{2} \int_\Omega (1 + \varepsilon \cos x_1)^3 \|\nabla v\|^2 \, \mathrm{d}x - \varepsilon k \int_\Omega \varepsilon \sin x_1 v \, \mathrm{d}x \quad \text{s.t. } v \geq 0 \text{ a.e. on } \Omega$$

with $\Omega = (0, 2\pi) \times (0, 2b)$, $b = L/(2R)$ and a constant k that depends on various physical parameters. As in [DMM04], we assume this constant to be equal to 1.

The results for the choice $\varepsilon = 0.1$ and $b = 5$ are given in Table 6.2. We also observe perfect level-independent convergence and see that the free form algorithm performs slightly worse in this example.

A Signorini problem

We next consider a problem of the class described in Section 5.5 that is the 3-D version of a test problem from [HW05]. A cube made from steel is transformed by a rigid displacement and traction forces act on the four side surfaces (cf. Figure 6.3). The cube has frictionless contact to a rigid foundation. The transformation is the solution of the problem

$$\min_{v \in H_D^1(\Omega)} \int_\Omega \left[\mu \varepsilon(v) : \varepsilon(v) + \frac{\lambda}{2} (\operatorname{div} v)^2 - f \cdot v \right] \mathrm{d}x - \int_{\Gamma_N} t \cdot v \, \mathrm{d}S(x)$$
$$\text{s.t. } v^T n \leq g \text{ on } \Gamma_C,$$

Levels	dof	ML	Smooth	f	f'	time (sec.)
5	63^2	3	4	8	7	0.1
6	127^2	3	4	8	7	0.2
7	255^2	3	4	8	7	0.5
8	511^2	3	4	8	7	1.5
9	1023^2	3	4	8	7	5.6
5	63^2	3	4	8	7	0.1
6	127^2	3	5	9	8	0.2
7	255^2	4	5	10	9	0.5
8	511^2	3	4	8	7	1.4
9	1023^2	4	4	9	8	6.3

Table 6.2.: Results for Journal-Bearing problem V-cycle (top) and free form (bottom)

where $\varepsilon(v)$ denotes the linearized strain tensor. The following configuration is used:

- Reference domain $\Omega := (0,1)^3 \subset \mathbb{R}^3$.
- Neumann and Dirichlet boundary conditions

$$\Gamma_N := \{0,1\} \times [0,1]^2 \cup [0,1] \times \{0,1\} \times [0,1],$$
$$\Gamma_D := [0,1]^2 \times \{1\}, \quad \Gamma_C := [0,1]^2 \times \{0\}.$$

- Displacements $H^1_D(\Omega) := \{u \in H^1(\Omega)^3 \mid u = (0,0,-0.07)^T \text{ on } \Gamma_D\}$.
- Material constants (steel): Shear modulus $\mu = E/(2+2\nu)$, Lame's first parameter $\lambda = E\nu/((1+\nu)(1-2\nu))$, Young modulus $E = 200$ and Poisson's ratio $\nu = 0.3$.
- Volume forces $f \equiv 0$ and boundary forces $t = (10(1-2x), 0, 6.5)^T$ on Γ_N.
- The gab g between the cube in reference configuration and the obstacle is 0.03.

We discretized the problem using piecewise linear continuous tetrahedron elements, the coarsest mesh consists of 27 nodes. In each node we have three degrees of freedom. Since the normal vector n on Γ_C is in every point equal to $(0,0,-1)^T$, we have simple bound constraints on the z component of the displacement in the discrete case. Hence, we do not need the special discretization basis of Section 5.5.1.

The results in the first half of Table 6.3 shows that the algorithm performs level-independently in this example. Even more, the number of iterations decreases, which is based on the fact that the initial value obtained by the nested iteration scheme becomes better as the number of levels increase. Without the full multigrid, the iteration number stays nearly constant (cf. bottom part of Table 6.3).

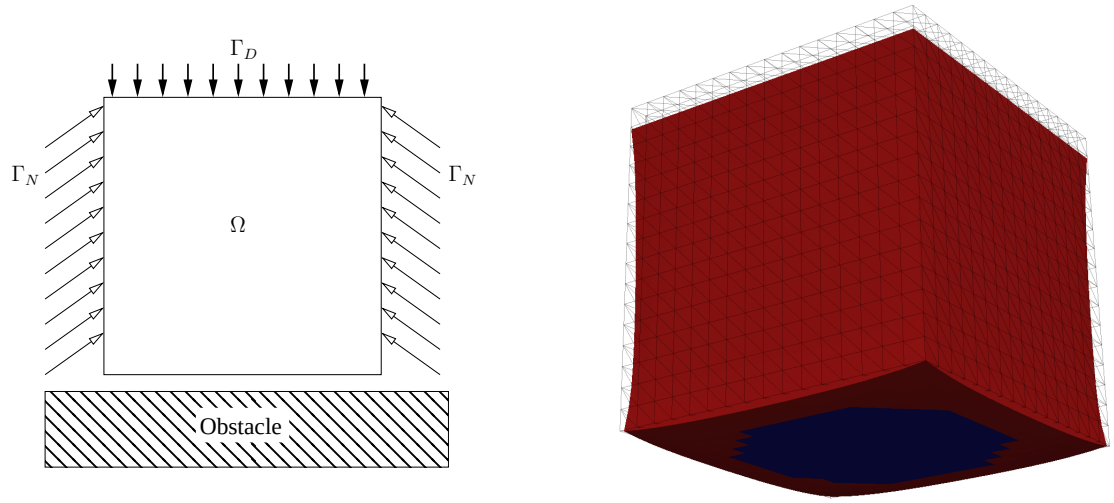

Figure 6.3.: Profile view of the cube and the rigid obstacle (left). Solution where the contact area is colored blue (right)

6.3.2. Minimum surface problems

The minimum surface problem, which we introduced in Section 5.4, has in comparison to the previous problems a non-quadratic objective function.

Enneper's Minimal Surface

We first let the algorithm determine Enneper's Minimal surface, which is a test problem from [ACM91]. It is the solution of the problem

$$\min_{u \in \mathcal{C}} \int_\Omega \sqrt{1 + \nabla u^T \nabla u}\, \mathrm{d}x \tag{6.2}$$

where $\Omega = (-1/2, 1/2) \times (-1/2, 1/2)$ and the convex set $\mathcal{C}$ is defined by

$$\mathcal{C} = \Big\{ u \in H^1(\Omega) \,|\, u(x) = u_D(x) \text{ for } x \in \partial\Omega \Big\}.$$

The boundary function $u_D \colon \mathbb{R}^2 \to \mathbb{R}$ is implicitly given as solution of $u_D(x) = v^2 - w^2$, where v and w are the unique solutions of the equations

$$x_1 = v + vw^2 - \frac{1}{3}v^3, \quad x_2 = -w - v^2 w + \frac{1}{3}w^3.$$

Because we directly incorporate the boundary condition into the discretization, this problem is unconstrained.

The results in Table 6.4 show again level-independent convergence behaviour.

Levels	dof	ML	Smooth	f	f'	time (sec.)
3	1,936	10	11	22	21	1.5
4	15,488	10	11	22	21	6.9
5	104,329	9	10	20	19	47.2
6	810,000	8	9	18	17	351.6
3	1,936	9	11	21	20	1.2
4	15,488	10	11	22	21	6.6
5	104,329	10	11	22	21	48.2
6	810,000	11	12	24	23	408.6

Table 6.3.: Results for the Signorini problem with (top) and without full multigrid strategy (bottom)

Levels	dof	ML	Smooth	f	f'	f''	time (sec.)
5	63^2	3	4	8	7	2	0.3
6	127^2	3	4	8	7	2	0.5
7	255^2	3	4	8	7	2	1.0
8	511^2	4	5	10	9	2	3.1
9	1023^2	4	5	10	9	1	10.3

Table 6.4.: Results for Enneper's Minimal Surface Problem

Minimum Surface with Obstacle

The next example, taken from [DMM04], is also a minimum surface problem but this time the surface is not determined by the boundary values alone but must also lie above an obstacle. We seek a solution of the problem (6.2) with $\Omega = (0,1)^2$ and where the feasible set is given by

$$\mathcal{C} = \Big\{ u \in H^1(\Omega) \,|\, u(x) = u_D(x) \text{ for } x \in \partial\Omega \text{ and } u(x) \geq l(x) \text{ a.e. on } \Omega \Big\}.$$

The boundary function is defined by

$$u_D(x) := \begin{cases} 1 - (2x_1 - 1)^2, & x_2 \in \{0, 1\}, \\ 0, & \text{otherwise}, \end{cases}$$

and the obstacle by

$$l(x) := \begin{cases} 1, & \text{if } |x_1 - 1/2| \leq 1/4,\ |x_2 - 1/2| \leq 1/4, \\ 0, & \text{otherwise}. \end{cases}$$

This problem does not possess a continuous solution which makes it difficult to solve. The slope of the discrete solutions near the obstacle goes to infinity as the mesh size approaches zero (cf. Figure 6.4). This difficulty was also observed for minimum surface problems on non-convex domains (cf., e.g., [Cia78, Ch. 5]).

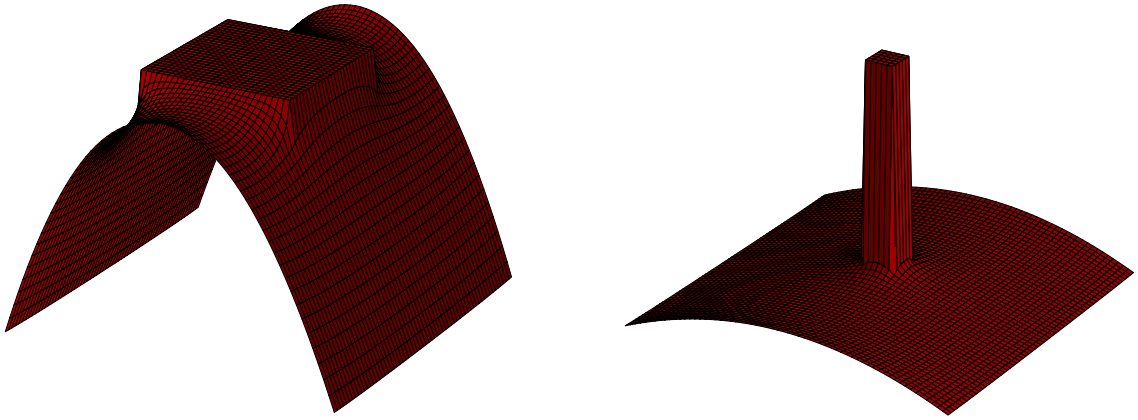

Figure 6.4.: Plots of the solutions of two minimum surface problems with obstacle (left: Example from [DMM04], right: Example from [GMS+10])

Remark 6.2 We will give a heuristic motivation why an infinite slope is very likely to result in numerical problems. Consider the family of 1-D functions $f_h\colon [0,1] \to \mathbb{R}$ defined by

$$f_h(x) := \begin{cases} 0, & \text{if } 0 \le x \le 1-h, \\ (x-(1-h))/h, & \text{otherwise.} \end{cases}$$

A simple calculation shows that the $H^1(\Omega)$-semi-norm of f_h goes to infinity as $h \to 0$. Even more, if we consider the sequence $(h_i)_{\mathbb{N}}$ with $h_i = 2^{-i}$, the distance between two elements is $|f_{h_i} - f_{h_{i-1}}|_{H^1(\Omega)} = |f_{h_{i-1}}|_{H^1(\Omega)}$. Hence, the distance goes to infinite as $i \to \infty$. The solutions to the discrete minimum surface problem with obstacle behave like the functions f_{h_i} near the obstacle when the meshsize of the finite element space is h_i. Although the full multigrid calculates an approximation of the solution on the space with mesh size h_{i-1}, it is hence no good initial value since its distance to the solution grows (in terms of $H^1(\Omega)$) as i goes to infinity. A standard multigrid method for linear, elliptic problems converges linear in terms of the energy norm, which is equivalent to the H^1-semi-norm. If we make the plausible assumption that our trust-region algorithm will at best converges like a standard multigrid method in this example, the number of steps on the fine grid also increases for larger i.

Hence, we would not expect that the algorithm performs level-independently, which is confirmed by the numerical results in the first part of Table 6.5.

The considerations in the previous remark suggests that the main errors occur near the obstacle. In order to obtain faster convergence, we made a slight modification to our smoother. For this we determine all non-active grid nodes which are near the active set. More precisely, given a fixed integer $l > 0$, the set $\mathcal{B}_l$ contains all non-active nodes which are connected by at most l edges to an active node. This set is then used to make additional smoothing sweeps using just the nodal basis functions corresponding to the nodes in $\mathcal{B}_l$. Normally, $\mathcal{B}_l$ is only a small subset of the complete set of nodes and the additional costs for the extra smoothing cycles are low. For the results in the second half of Table 6.5 we have set $l = 5$ and made six additional smoothing sweeps on $\mathcal{B}_l$ after every full smoothing cycle. We see a much better performance of this variant and a weaker dependence on the number of levels used.

Levels	dof	ML	Smooth	f	f'	f''	time (sec.)
5	63^2	8	15	24	23	3	0.6
6	127^2	10	21	32	30	3	1.3
7	255^2	15	24	40	38	3	3.8
8	511^2	19	34	54	51	4	16.5
9	1023^2	28	44	73	70	3	85.6
5	63^2	5	10	16	14	3	0.4
6	127^2	6	12	19	17	3	0.9
7	255^2	9	14	24	20	3	2.7
8	511^2	8	17	26	23	3	9.7
9	1023^2	12	19	32	28	4	47.5

Table 6.5.: Results for Minimum Surface problem with obstacle (top) and with additional smoothing steps near the active set (bottom)

Levels	dof	ML	Smooth	f	f'	f''	time (sec.)
5	63^2	5	19	25	19	7	0.6
6	127^2	6	20	27	20	7	1.2
7	255^2	7	26	34	22	8	3.7
8	511^2	22	39	62	48	8	20.5
9	1023^2	28	64	93	73	11	120.7

Table 6.6.: Results for minimum surface problem from [GMS+10] with additional smoothing steps near the active set

A similar example was considered in [GMS+10]. Here, the feasible set is given by

$$\mathcal{C} = \left\{ u \in H^1(\Omega) \,|\, u(x) = u_D(x) \text{ for } c \in \partial\Omega \text{ and } u(x) \geq l(x) \text{ a.e. on } \Omega \right\}$$

with

$$u_D(x) := \begin{cases} x_1(1-x_1), & x_2 \in \{0,1\}, \\ 0, & \text{otherwise}, \end{cases}$$

and the obstacle

$$l(x) := \begin{cases} \sqrt{2}, & \text{if } |x_1 - 1/2| \leq 1/18,\ |x_2 - 1/2| \leq 1/18, \\ 0, & \text{otherwise}. \end{cases}$$

This one is even more difficult to solve since the obstacle is higher than in the previous example (cf. Figure 6.4), which is confirmed by the results in Table 6.6.

In a final minimum surface example we now will show the positive effect of the active-set strategy. To this end, we use the same data as in the last problem except for the lower bound which we set

Levels	dof	ML	Smooth	f	f'	f''	time (sec.)
5	63^2	4	6	11	10	2	0.3
6	127^2	4	6	11	10	2	0.5
7	255^2	4	6	11	10	2	1.1
8	511^2	4	6	11	10	2	3.7
9	1023^2	4	6	11	10	2	14.6
5	63^2	26	27	54	53	2	1.0
6	127^2	33	34	68	67	2	2.4
7	255^2	41	42	84	83	2	6.9
8	511^2	47	48	96	95	2	25.6
9	1023^2	53	54	108	107	2	110.9

Table 6.7.: Results for Minimum Surface problem with single point obstacle with (top) and without the active set strategy (bottom)

to

$$l(x) := \begin{cases} \sqrt{2}, & \text{if } x_1 = x_2 = 1/2 - h_r, \\ 0, & \text{otherwise,} \end{cases}$$

where h_r denotes the grid size of the finest mesh. In the solution, the function is active at exactly one fine grid node. As discussed in Section 4.3.2, in this setting the lower-level steps are zero near the active point if we do not use the active set strategy.

As we can see in Table 6.7, the differences are huge. With the active set strategy the algorithms converges in a level independent number of steps, whereas without the active set strategy we observe a dependence of the size $O(\log h_r)$.

6.3.3. Example on a non-convex domain

All domains in the previous examples were convex, and in this case we have a strong regularity result for second-order elliptic PDEs. Thus, we consider in the next example an L-shaped domain with reentrant corner (Figure 6.5).

On this domain, we solve the following problem whose objective function is non-convex and of the type discussed in Section 5.3:

$$\min_{u \in \mathcal{H}^1_D(\Omega)} \frac{1}{2} \int_\Omega \Big[\Big(3x_1 \sin(4\pi u)^2 + \frac{1}{4} \Big) \nabla u^T \nabla u + u^2 \Big] \, \mathrm{d}x$$

where

$$H^1_D(\Omega) := \Big\{ u \in H^1(\Omega) \,|\, u(x) = 0 \text{ for } x \in \Gamma_0,\ u(x) = 1 \text{ for } x \in \Gamma_1 \Big\}.$$

The results in Table 6.8 show again nearly perfect level independent convergence for this example.

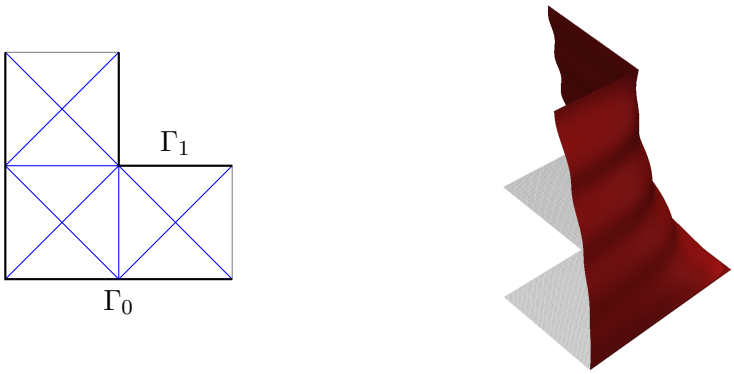

Figure 6.5.: Domain Ω with initial triangulation (left) and plot of the solution (right)

Levels	dof	ML	Smooth	f	f'	f''	time (sec.)
7	24,447	19	20	40	38	3	5.0
8	98,047	21	23	45	44	2	16.3
9	392,703	20	21	42	41	2	56.6
10	1,571,839	19	20	40	39	1	214.3

Table 6.8.: Results for problem on domain with reentrant corner

6.3.4. Optimal design with composite materials

Another problem from [ACM91] requires determining the placement of two elastic materials in the cross-section of a rod with maximal torsional rigidity. We will not go further in the details of the modeling. The problem to solve is given by

$$\min_{u \in H_0^1(\Omega)} \int_\Omega (\psi_\lambda(\|\nabla u\|) + u)\,\mathrm{d}x$$

where $\psi_\lambda \colon \mathbb{R} \to \mathbb{R}$ is a piecewise quadratic function defined by

$$\psi_\lambda(t) := \begin{cases} \frac{1}{2}\mu_2 t^2, & 0 \le t \le t_1, \\ \mu_2 t_1 (t - \frac{1}{2} t_1), & t_1 < t \le t_2, \\ \frac{1}{2}\mu_1 (t^2 - t_2^2) + \mu_2 t_1 (t_2 - \frac{1}{2} t_1), & t_2 < t, \end{cases}$$

with the breakpoints

$$t_1 = \sqrt{2\lambda \frac{\mu_1}{\mu_2}} \quad \text{and} \quad t_2 = \sqrt{2\lambda \frac{\mu_2}{\mu_1}}.$$

Here, $\Omega = (0,1)^2$ and the parameters are $\lambda = 0.008$, $\mu_1 = 1$ and $\mu_2 = 2$.

The main difficulty of this problem lies in the fact that the function ψ_λ is not twice continuously differentiable and hence the whole functional does not satisfy our differentiability assumptions. Nonetheless, the standard V-cycle version of our algorithm worked very well but failed to converge in a reasonable time on the finest level. Hence, we also tried it with the free-form version which was capable to solve also the fine-level problem (cf. Table 6.9).

Levels	dof	ML	Smooth	f	f'	f''	time (sec.)
6	16,129	26	26	53	44	10	4.9
7	65,025	37	31	69	54	16	14.8
8	261,121	87	78	166	135	45	122.0
9	1,046,529	–	–	>1000	–	–	–
6	16,129	41	28	70	56	8	8.8
7	65,025	21	27	49	40	18	12.9
8	261,121	42	44	87	75	19	87.6
9	1,046,529	24	38	63	55	23	249.9

Table 6.9.: Results for optimal design problem, V-cycle (top), free form (bottom)

6.3.5. Nonlinear elasticity

Our final set of test problems consists of finding deformations of bodies made from hyperelastic material. The problem class was already presented in Section 5.6. The solution images show the deformed bodies, where the displacement vectors are not amplified.

Twisting of a hyperelastic cube

In the first example, we consider a cube whose bottom side is clamped to a fixed foundation. The top surface is rotated by 60 degrees and no forces operate on the cube, which is assumed to be made of a compressible Mooney-Rivlin material with Young modulus $E = 200$ and Poisson's ration $\nu = 0.3$. The results in Table 6.10 show again level-independent convergence of the algorithm in this example.

In comparison to the previous problems, the objective function is much more difficult to evaluate. As an example, the calculation of the Hessian on the finest level takes roughly 140 seconds. The proportion between the time we spend on evaluating the function and the time we use for the smoothing is not balanced very well. Hence, one can hope for faster convergence by making more smoothing cycles in one smoothing step. This guess is approved by the results in Table 6.10 where we compare our standard choice $m = 6$ against $m = 60$. The difference is even larger in the next example.

A buckling plate

A typical phenomena which is observed in reality is buckling of an elastic body. Buckling occurs if compressive stress is so large that the body buckles in one direction to reduce its stress. For example, consider a piece of paper that is held tight between two hands. If the hands move together, the paper "buckles" in one direction that is perpendicular to the movement direction of the hands. In general, the final state is not unique as there are two directions in which the paper could buckle. With this in mind, it is obvious that this effect cannot be observed in

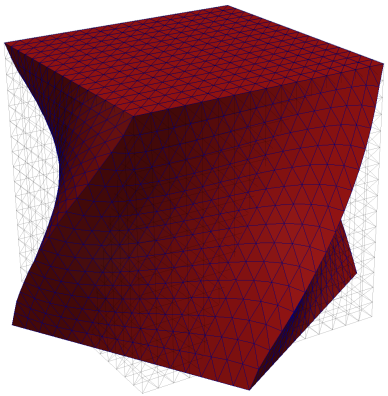

Figure 6.6.: Solution of twisted cube example (4 levels)

Levels	dof	ML	Smooth	f	f'	f''	time (sec.)
3	1,681	7	10	18	17	4	5.9
4	12,996	8	10	19	18	3	19
5	101,124	8	10	19	18	3	112.9
6	797,449	7	8	16	15	2	661.8
3	1,681	1	4	6	5	3	5.1
4	12,996	3	5	9	8	3	16.3
5	101,124	3	5	9	8	2	79.5
6	797,449	3	5	9	8	2	629.0

Table 6.10.: Results of the twisted cube example with $m = 6$ (top) and $m = 60$ (bottom) cycles in the smoothing algorithm

simulations that use a linear elasticity model, since the solution of these are unique (cf. [Cia88, Theorem 6.3-5]).

The basic configuration of our problem is the following: A plate made of a hyperelastic compressible Mooney-Rivlin material, which we already introduced in Section 5.6, is clamped to a wall on the face Γ_D (cf. Figure 6.7). The face Γ'_D undergoes a rigid translation in direction of the first unit vector e_1. More precisely, we seek a displacement u^* that solves

$$\min_{u \in H^1_D(\Omega,\mathbb{R}^3)} \int_\Omega \Big(a\|F\|^2 + b\|\operatorname{cof} F\|^2 + \gamma(\det F) - (3a+3b+c) \Big)\, \mathrm{d}x, \quad F(x) := I + \nabla u(x),$$

with

$$H^1_D(\Omega,\mathbb{R}^3) := \{u \in H^1(\Omega,\mathbb{R}^3) \,|\, u(x) = 0 \text{ on } \Gamma_D, \quad u(x) = -\frac{x_3}{2} e_1 \text{ on } \Gamma'_D\},$$

$\gamma(\delta) := c\delta^2 - d\log(\delta)$ and the parameters

$$a = \frac{\mu}{2} + c - \frac{\lambda}{4}, \quad b = \frac{\lambda}{4} - c, \quad c = \frac{\lambda}{5}, \quad d = \frac{\lambda}{2} + \mu.$$

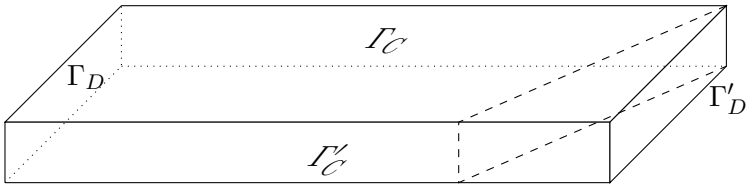

Figure 6.7.: Reference configuration Ω of the buckling plate example

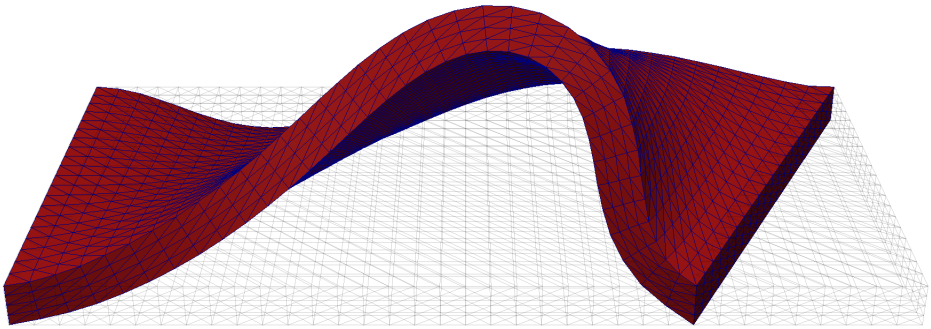

Figure 6.8.: Solution of the unconstrained buckling plate problem with 3 Levels

The Lamï¿½ parameters λ and μ are given by

$$\lambda = \frac{E\nu}{(1-2\nu)(1+\nu)}, \quad \mu = \frac{E}{2(1+\nu)}$$

with Young's modulus $E = 200$ and Poisson's parameter $\nu = 0.3$.

As in the last example, we increase the number of cycles in one smoothing iteration to $m = 60$.

We first solve the problem without any additional constraints (cf. Figure 6.8). The results in Table 6.11 show that the algorithm is quite fast and the time grows linearly with the degrees of freedom. The second part of the table shows the result of the free form variant, which clearly outperforms the V-cycle algorithm. The algorithm uses the lower levels more intensively and the form is more like a W-cycle.

As a last example, we add additional constraints to the feasible displacement in direction e_2 on the top and bottom surface and assume that the contact between the body and the rigid obstacle is frictionless. We seek a solution to

$$\min_{u\in\mathcal{C}} \int_\Omega \Big(a\|F\|^2 + b\|\operatorname{cof} F\|^2 + \gamma(\det F) - (3a+3b+c)\Big)\,\mathrm{d}x, \quad F(x) := I + \nabla u(x),$$

with the feasible set

$$\mathcal{C} = \{u \in H^1_D(\Omega,\mathbb{R}^3) \,|\, u_2(x) \le 0.2 \text{ on } \Gamma_C, \quad u_2(x) \ge -0.2 \text{ on } \Gamma'_C\}.$$

Levels	dof	ML	Smooth	f	f'	f''	time (sec.)
3	11,025	82	84	167	166	5	41.0
4	78,597	68	69	138	137	2	190.9
5	590,733	64	65	130	129	1	1172.66
3	11,025	53	59	113	112	7	40.7
4	78,597	20	24	45	44	3	135.8
5	590,733	11	14	26	25	2	519.3

Table 6.11.: Results of the unconstrained buckling plate problem with V-cycle version (top) and free form version (bottom)

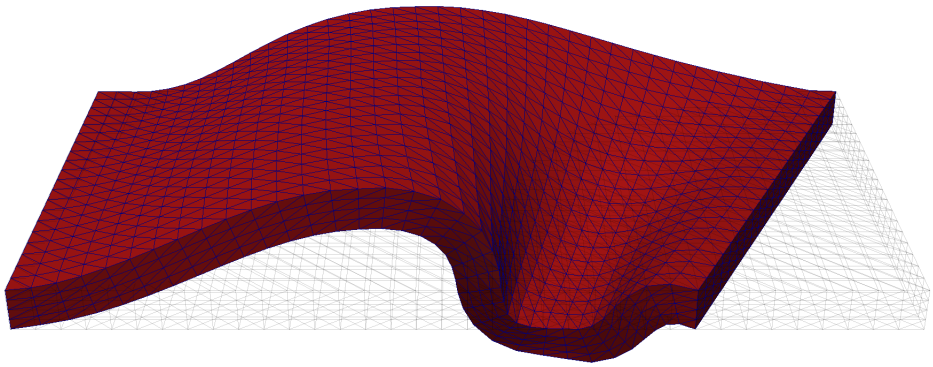

Figure 6.9.: Solution of the constrained buckling plate problem with 3 Levels

Levels	dof	ML	Smooth	f	f'	f''	time (sec.)
3	11,025	143	122	266	241	11	64.0
4	78,597	147	149	297	296	4	389.6
5	590,733	128	128	257	255	2	2413.5
3	11,025	146	125	272	245	12	94.7
4	78,597	69	68	138	131	6	326.4
5	590,733	32	36	69	68	3	1199.3

Table 6.12.: Results of the constrained buckling plate problem with V-cycle version (top) and free form version (bottom)

Figure 6.9 shows the final solution. We see that the constraints at the bottom surface is active in some points and the solution buckles twice.

Similar to the unconstrained example, the free form algorithm perform much better on five levels (cf. Table 6.12).

A. Appendix

In this appendix we summarize some result from functional analysis which we often use in the preceding chapters. Furthermore, we show a differentiability result for nonlinear variational problems.

A.1. Sobolev embeddings

Theorem A.1 *Let $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, be a bounded domain with Lipschitz-boundary and furthermore $m_1 \geq m_2 \geq 0$. Then the embedding*

$$W^{m_1,p_1}(\Omega) \hookrightarrow W^{m_2,p_2}(\Omega)$$

exists and is continuous if $m_1 - n/p_1 \geq m_2 - n/p_2$. In this cases, the following inequality is satisfied:

$$\|u\|_{W^{m_2,p_1}(\Omega)} \leq C\|u\|_{W^{m_1,p_2}(\Omega)}.$$

The embedding is compact if $m_1 - n/p_1 > m_2 - n/p_2$.

Proof See for example [Alt06, Thm. 8.9]. □

A.2. Projections in Hilbert spaces

Theorem A.2 (Projection Theorem) *Let U be a Hilbert space and $\emptyset \neq C \subset U$ a closed and convex set. Then there exists a unique mapping $Proj_C \colon U \to C$ that satisfies*

$$\|x - Proj_C(x)\|_U = \inf_{y \in C} \|x - y\|_U \quad \text{for all } x \in U.$$

Furthermore, the projection on the set C can also be defined as the unique operator that fulfills for every $x \in U$:

$$(x - Proj_C(x), Proj_C(x) - y)_U \geq 0 \quad \text{for all } y \in C.$$

If C is a subspace, then it holds:

$$(x, y)_U = (Proj_C(x), y)_U \quad \text{for all } y \in C.$$

Proof See, e.g., [HPUU09, Lemma 1.10]. □

The next lemma shows that the projection operator is Lipschitz continuous with constant $L = 1$ and monotone.

Lemma A.1 *Let U be a Hilbert space and $\emptyset \neq C \subset U$ a closed and convex set. Then the projection satisfies*

$$\|Proj_C(x) - Proj_C(y)\|_U \leq \|x - y\|_U \quad \text{for all } x, y \in U$$

and

$$(x - y, Proj_C(x) - Proj_C(y))_U \geq 0 \quad \text{for all } x, y \in U.$$

Proof See, e.g., [HPUU09, Lemma 1.10].

A.3. Weak convergence

Definition A.1 Let V be a normed space with dual space V^*. A sequence $(v_k) \subset V$ is said to *converge weakly* to an element v ($v_k \rightharpoonup v$) if

$$\langle f, v_k \rangle \to \langle f, v \rangle \quad \text{for all } f \in V^*.$$

Theorem A.3 (Eberlein-Shmulyan) *A Banach space V is reflexive iff every strongly bounded sequence of V contains a subsequence which converges weakly to an element of V.*

Proof See, e.g., [Yos80, Section V.4]. □

Lemma A.2 *Let V be a normed vector space. If $C \subset V$ is closed and convex, then it is* weakly sequentially closed*, i.e., for every weakly convergent sequence $(v_k) \subset C$ with $v_k \rightharpoonup v$ also $v \in C$ is satisfied.*

Proof See, e.g., [Alt06]. □

A.4. Differentiability in Banach spaces

In this section we summarize some basic results about differentiability in Banach spaces, a more extensive presentation can be found, e.g., in [IT79].

In the following, let X and Y be Banach spaces and U be an open subset of X. A function $f\colon U \to Y$ is said to be *Gâteaux differentiable at $x \in U$* if the limit

$$\lim_{t \to 0} t^{-1}(f(x + ts) - f(x)) =: f'(x)[s]$$

exists for all directions $s \in X$ and the mapping $s \mapsto f'(x)[s]$ is linear and continuous. If this holds for all $x \in U$, we call the mapping f *Gâteaux differentiable*. If $Y = \mathbb{R}$, then $f'(x) \in X^*$ and we will also use the dual pair notation $\langle f'(x), s \rangle$ for $f'(x)[s]$.

If there exists a bounded linear operator $L_x \in \mathcal{L}(X, Y)$ such that

$$f(x+s) = f(x) + L_x s + \phi(s), \quad \|\phi(s)\|_Y / \|s\|_X \to 0 \text{ as } \|s\|_X \to 0,$$

for all s with $x + s \in U$, the function is called *Fréchet differentiable at* $x \in U$. One can show that in this case f is continuous and also differentiable in the Gâteaux sense with $f'(x)[s] = L_x s$.

If f is Gâteaux or Fréchet differentiable and the mapping $x \mapsto f'(x)$ is continuous on the Banach space $\mathcal{L}(X, Y)$, then f is said to be *continuously differentiable*. One can show that in this setting a continuously Gâteaux differentiable function is also Fréchet differentiable. Hence, it will cause no confusion if we do not distinguish between Gâteaux and Fréchet continuously differentiability in this case.

A function $f\colon U \to Y$ is twice Gâteaux (Fréchet) differentiable if f and its first derivative $f'\colon U \to \mathcal{L}(X, Y)$ are Gâteaux (Fréchet) differentiable. Similar, higher-order differentiability is defined inductively. One can show (cf. [Zei86, Prop. 4.20]) that f has a n-th Gâteaux derivative at $x \in U$ iff f has a $(n-1)$th Gâteaux derivative at x and

$$f^{(n)}(x)[s_1, \ldots, s_n] := \lim_{t \to 0^+} \frac{f^{(n-1)}(x + t s_n)[s_1, \ldots, s_{n-1}] - f^{(n-1)}(x)[s_1, \ldots, s_{n-1}]}{t}$$

exists, is n-linear and bounded.

We need the following generalized version of Taylor's theorem:

Lemma A.3 (Taylor's Theorem) *Let $f\colon U \to \mathbb{R}$ be n-times Gâteaux differentiable at every point of the interval $[x, x+s] \subset U$ and let the mapping $x \mapsto f^{(n)}(x)[s, \ldots, s]$ be continuous. Then*

$$f(x+s) = f(x) + f'(x)[s] + \frac{1}{2!} f''(x)[s, s] + \ldots + \frac{1}{(n-1)!} f^{(n-1)}(x)[s, \ldots, s] + R_n(x)$$

with

$$R_n(x) := \frac{1}{(n-1)!} \int_0^1 (1-t)^{n-1} f^{(n)}(x + ts)[s, s, \ldots, s] \, \mathrm{d}t.$$

Proof [Zei86, Theorem 4.A] □

Lemma A.4 *Let $f\colon U \subset X \to Y$ be a Gâteaux differentiable function in a neighborhood $U(x)$ of x, then for all $s \in X$ with $\{x + ts \,|\, t \in [0,1]\} \subset U(x)$, the following holds:*

$$\|f(x+s) - f(x)\|_Y \leq \sup_{0 \leq t \leq 1} \|f'(x + ts)[s]\|_Y.$$

If there exists a constant L such that $\|f'(\bar{x})\|_{\mathcal{L}(X,Y)} \leq L$ for all $\bar{x} \in U$, then f is Lipschitz continuous with Lipschitz constant L.

Proof [HPUU09, Section 1.4.1] □

A.4.1. Differentiability of variational integrals

Let $\Omega \subset \mathbb{R}^d$ be a nonempty, open and bounded measurable set. We are interested in the differentiability of the functional

$$J(u) := \int_\Omega f(x, u, \nabla u)\,\mathrm{d}x, \tag{A.1}$$

where $f\colon \Omega \times \mathbb{R}^N \times \mathbb{R}^{dN}$, $(x, u, z) \mapsto f(x, u, z)$, is a Carathéodory function, that means f is measurable in x for each $(u, z) \in \mathbb{R}^N \times \mathbb{R}^{dN}$ and continuous in (u, z) for almost all $x \in \Omega$.

Before we start, we shortly introduce vector-valued Lebesgue- and Sobolev-spaces.

Definition A.2 Let $\Omega \subset \mathbb{R}^d$ be a domain, $1 \le p \le \infty$, $N \in \mathbb{N}$ and $m \ge 0$. The space $W^{m,p}(\Omega)^N$ consists of all functions $u\colon \Omega \to \mathbb{R}^N$ with $u_i \in W^{m,p}(\Omega)$ for all $i = 1, \dots, N$.

Remark A.1 Normally, if $u \in W^{m,p}(\Omega)^N$, $m \ge 1$, the weak gradient $\nabla u(x)$ is a matrix with dimensions $d \times N$. In this section, however, we will identify it instead with a vector in $\mathbb{R}^{dN}$.

Lemma A.5 *Let $\|\cdot\|$ be an arbitrary norm on $\mathbb{R}^N$. If $u \in L^p(\Omega)^N$, then $\|u\| \in L^p(\Omega)$.*

Proof We use the fact that all norms on $\mathbb{R}^N$ are equivalent. Hence, with a constant C_N that does only depend on N we obtain:

$$\int_\Omega \|u(x)\|^p\,\mathrm{d}x \le C_N^p \int_\Omega \Big(\sum_{i=1}^N |u_i(x)|^p\Big)\,\mathrm{d}x = C_N^p \sum_{i=1}^N \|u_i\|_{L^p(\Omega)}^p. \qquad \square$$

In the following we denote by f_u resp. f_z the derivative of f with respect to u resp z. It is well-known that functions of the type (A.1) are continuously differentiable under certain growth assumptions:

Theorem A.4 *Assume that with a constant $C > 0$ either*

$$\begin{aligned} &|f(x, u, z)| \le CV^2, \\ &\|f_u(x, u, z)\| \le CV,\ \|f_z(x, u, z)\| \le CV, \\ &V := (1 + \|u\|^2 + \|z\|^2)^{1/2}. \end{aligned} \tag{A.2}$$

is satisfied for all $(u, z) \in \mathbb{R}^N \times \mathbb{R}^{dN}$ and almost all $x \in \Omega$, or for all $R > 0$, $\|u\| < R$, $z \in \mathbb{R}^{dN}$ and almost all $x \in \Omega$

$$\begin{aligned} &|f(x, u, z)| \le C(R)\tilde{V}^2, \\ &\|f_u(x, u, z)\| \le C(R)\tilde{V}^2,\ \|f_z(x, u, z)\| \le C(R)\tilde{V}, \\ &\tilde{V} := (1 + \|z\|^2)^{1/2}. \end{aligned} \tag{A.3}$$

holds with a constant $C(R) > 0$ that does depend on R. Then:

1. *If assumptions* (A.2) *are satisfied, J is continuously differentiable on $H^1(\Omega)^N$.*

2. *If instead* (A.3) *hold, J is continuously differentiable on $H^1(\Omega)^N \cap L^\infty(\Omega)^N$, the norm on this space being*

$$\|u\|_{H^1(\Omega)^N \cap L^\infty(\Omega)^N} := \|u\|_{L^\infty(\Omega)^N} + \|\nabla u\|_{L^2(\Omega)^{dN}}.$$

In both cases the derivative in direction d is given by

$$J'(u)[d] = \int_\Omega \Big(f_u(x,u,\nabla u)d + f_z(x,u,\nabla u)\nabla d \Big)\,\mathrm{d}x.$$

Proof [Mor08, Theorem 1.10.3]. □

Remark A.2 Assumptions (A.2) can be weakened by considering the embeddings given by Theorem A.1. For this we need the additional assumption that Ω has a Lipschitz-continuous boundary. For example, in the case $d = 2$, the following conditions, $2 \le q < \infty$ arbitrary, can be used instead of (A.2):

$$\begin{aligned} |f(x,u,z)| &\le C(g_1(x) + \|u\|^q + \|z\|^2), && g_1 \in L^1(\Omega) \\ \|f_u(x,u,z)\| &\le C(g_2(x) + \|u\|^{q-1} + \|z\|^{2-2/q}), && g_2 \in L^{q/(q-1)}(\Omega) \\ \|f_z(x,u,z)\| &\le C(g_3(x) + \|u\|^{q/2} + \|z\|), && g_3 \in L^2(\Omega). \end{aligned} \tag{A.4}$$

The situation is more complicated for the second derivative of J. Even under restrictive growth conditions, functions of the type (A.1) are in general neither twice continuously nor Fréchet differentiable. An example can be found in [Nol93]. Instead, we will show that J is twice Gâteaux differentiable and that the mapping $u \mapsto J''(u)[d,d]$ is continuous for each direction d. To this end, we need the following two preliminary lemmas.

Lemma A.6 *Let $\Omega \subset \mathbb{R}^d$ be a nonempty measurable set and $\phi\colon \Omega \times \mathbb{R}^k \to \mathbb{R}$ be a Carathéodory function, i.e., ϕ is measurable for all $x \in \Omega$ and continuous in $v \in \mathbb{R}^k$ for almost all x. Suppose that the following growth condition with $g \in L^q(\Omega)$, $b \ge 0$ and $1 \le p, q < \infty$ is satisfied:*

$$|\phi(x,v)| \le g(x) + b\|v\|^{p/q}.$$

Then $\Phi\colon L^p(\Omega)^k \to L^q(\Omega)$ with $\Phi(v)(x) := \phi(x, v(x))$ is continuous and bounded with

$$\|\Phi(v)\|_{L^q(\Omega)} \le C(\|g\|_{L^q(\Omega)} + \|v\|^{p/q}_{L^p(\Omega)^k}).$$

Proof [Zei90, Theorem 26.6] □

Lemma A.7 *Let $\Omega \subset \mathbb{R}^d$ be bounded. Furthermore let $\phi\colon \Omega \times \mathbb{R}^N \to \mathbb{R}^{m_1 \times m_2}$ be a bounded Carathéodory function, i.e., there exists a constant M such that $\|\phi(x,u)\| \le M$ for all u and almost all $x \in \Omega$. Then*

$$G(u) := \int_\Omega d_1(x)^T \phi(x,u) d_2(x)\,\mathrm{d}x$$

with $d_1 \in L^2(\Omega)^{m_1}$ and $d_2 \in L^2(\Omega)^{m_2}$ is continuous on $L^2(\Omega)^N$.

PROOF It is clear that $G(u) < \infty$ because of the boundedness of ϕ.

Let $(u_k) \subset L^2(\Omega)^k$ be a sequence that converges to u. We assume that $G(u_k) \not\to G(u)$. Then there exists a subsequence $(u_{\hat{k}})$ such that $|G(u_{\hat{k}}) - G(u)| \geq \delta > 0$ for all $\hat{k}$.

The function ϕ is bounded and hence satisfies the growth condition of Lemma A.6 for arbitrary $1 \leq q < \infty$ (componentwise). Therefore, the operator $\Phi\colon L^2(\Omega)^k \to L^q(\Omega)^{m_1 \times m_2}$, $\Phi(u)(x) := \phi(x, u(x))$, is continuous. Setting $\Phi_{\hat{k}} := \Phi(u_{\hat{k}})$ and $\Phi_* := \Phi(u)$ we have $\Phi_{\hat{k}} \to \Phi_*$ as $\hat{k} \to \infty$ in $L^q(\Omega)$. As a consequence, there exists a subsequence $(\Phi_{k'})$ of $(\Phi_{\hat{k}})$ with $\Phi_{k'}(x) \to \Phi_*(x)$ almost everywhere on Ω. Moreover, by Egorov's Theorem it follows that for each $\varepsilon > 0$ there exists a measurable set $E_\varepsilon \subset \Omega$ with $|\Omega \setminus E_\varepsilon| \leq \varepsilon$ and $\sup_{x \in E_\varepsilon} \|\Phi_{k'}(x) - \Phi_*(x)\| \to 0$ as $k' \to \infty$.

Using the boundedness of ϕ, we obtain for each $\varepsilon > 0$

$$\begin{aligned} |G(u_{k'}) - G(u)| &\leq \int_\Omega \|\phi(x, u_{k'}(x)) - \phi(x, u(x))\| \|d_1\| \|d_2\| \,\mathrm{d}x \\ &\leq \sup_{x \in E_\varepsilon} \|\phi(x, u_{k'}(x)) - \phi(x, u(x))\| \int_{E_\varepsilon} \|d_1\| \|d_2\| \,\mathrm{d}x + 2M \int_{\Omega \setminus E_\varepsilon} \|d_1\| \|d_2\| \,\mathrm{d}x. \end{aligned}$$

Since $\int_{\Omega \setminus E_\varepsilon} \|d_1\| \|d_2\| \,\mathrm{d}x \to 0$ as $\varepsilon \to 0$, we find an $\varepsilon^* > 0$ such that

$$2M \int_{\Omega \setminus E_{\varepsilon^*}} \|d_1\| \|d_2\| \,\mathrm{d}x < \frac{\delta}{2}.$$

Because Φ_k converges uniformly on the set E_{ε^*}, we find $K_{\varepsilon^*} \in \mathbb{N}$ such that

$$\sup_{x \in E_{\varepsilon^*}} \|\phi(x, u_{k'}(x)) - \phi(x, u(x))\| \int_{E_{\varepsilon^*}} \|d_1\| \|d_2\| \,\mathrm{d}x < \frac{\delta}{2} \quad \text{for } k' \geq K_{\varepsilon^*}.$$

This shows $|G(u_{k'}) - G(u)| < \delta$ for all $k' \geq K_{\varepsilon^*}$, which contradicts our assumption. Finally, this shows the continuity of G. □

Theorem A.5 *Suppose that the assumptions of Theorem A.4 hold. Moreover, let the function $(u, z) \mapsto f(x, u, z)$ be twice continuously differentiable for almost all $x \in \Omega$ and let either*

$$\|f_{uu}(x, u, z)\|, \|f_{zu}(x, u, z)\|, \|f_{zz}(x, u, z)\| \leq C \tag{A.5}$$

hold, or for all $R \geq 0$ and $\|u\| < R$

$$\begin{aligned} \|f_{uu}(x, u, z)\| &\leq C(R)\tilde{V}^2, \\ \|f_{uz}(x, u, z)\| &\leq C(R)\tilde{V}, \\ \|f_{zz}(x, u, z)\| &\leq C(R) \end{aligned} \tag{A.6}$$

with a constant $C(R) > 0$ and $\tilde{V}$ as in Theorem A.4 be satisfied.

1. *If* (A.2) *and* (A.5) *are satisfied, J is twice Gâteaux differentiable on $H^1(\Omega)^N$, and the operator $u \mapsto J''(u)[d, d]$ is continuous for every fixed direction $d \in H^1(\Omega)^N$.*
2. *If instead* (A.3) *and* (A.6) *hold, J is twice Gâteaux differentiable on $H^1(\Omega)^N \cap L^\infty(\Omega)^N$, and the operator $u \mapsto J''(u)[d, d]$ is continuous for every fixed direction $d \in H^1(\Omega)^N \cap L^\infty(\Omega)^N$.*

In both cases, the derivative in direction d_1, d_2 is given by

$$J''(u)[d_1,d_2] = \int_\Omega \Big(f_{uu}(x,u,\nabla u)[d_1,d_2] + f_{uz}(x,u,\nabla u)[d_1,\nabla d_2] \\ + f_{uz}(x,u,\nabla u)[\nabla d_1,d_2] + f_{zz}(x,u,\nabla u)[\nabla d_1,\nabla d_2]\Big)\,\mathrm{d}x,$$

and $J''(u)[d_1,d_2] = J''(u)[d_2,d_1]$ holds.

Proof It follows from Theorem A.4 that the functional J is continuously differentiable. We define $\varphi(t) := J'(u+td_2)[d_1]$ and formally differentiate φ:

$$\begin{aligned}
\varphi'(t) &= \frac{d}{dt} J'(u+td_2)[d_1] \\
&= \int_\Omega \frac{d}{dt}\Big(f_u(x,u+td_2,\nabla u + t\nabla d_2)[d_1] + f_z(x,u+td_2,\nabla u+t\nabla d_2)[\nabla d_1]\Big)\,\mathrm{d}x \\
&= \int_\Omega \Big(f_{uu}(x,u+td_2,\nabla u+t\nabla d_2)[d_1,d_2] + f_{uz}(x,u+td_2,\nabla u+t\nabla d_2)[d_1,\nabla d_2] \\
&\quad + f_{uz}(x,u+td_2,\nabla u+t\nabla d_2)[\nabla d_1,d_2] + f_{zz}(x,u+td_2,\nabla u+t\nabla d_2)[\nabla d_1,\nabla d_2]\Big)\,\mathrm{d}x \\
&=: \int_\Omega g(x,u,d_1,d_2,t)\,\mathrm{d}x.
\end{aligned}$$

To justify this formal argument, we have to show that the integrand is uniformly bounded by an integrable function in a neighborhood of $t=0$. Then we are allowed to interchange integration and differentiation.

We first assume that (A.5) holds. In this case, the integrand can be estimated independently of t by

$$|g(x,u,d_1,d_2,t)| \le C\Big[\|d_1(x)\|\big(\|d_2(x)\| + \|\nabla d_2(x)\|\big) + \|\nabla d_1(x)\|\big(\|d_2(x)\| + \|\nabla d_2(x)\|\big)\Big], \quad \text{(A.7)}$$

which is integrable since $\|d_1\|, \|d_2\|, \|\nabla d_1\|, \|\nabla d_2\| \in L^2(\Omega)$. Thus the directional derivative is given by $J''(u)[d_1,d_2] = \varphi'(0)$. Since f is twice continuously differentiable, $J''(u)[d_1,d_2]$ is linear in d_1 and d_2, and $J''(u)[d_1,d_2] = J''(u)[d_2,d_1]$ holds. Even more, from (A.7) follows (with a different constant C) that

$$J''(u)[d_1,d_2] \le C\|d_1\|_{H^1(\Omega)^N}\|d_2\|_{H^1(\Omega)^N},$$

which shows the boundedness of the differential $J''(u)$. Together with the linearity this implies the continuity of $J''(u)$ with respect to the directions. Therefore, J is twice Gâteaux differentiable.

Now assume that the weaker conditions (A.6) are satisfied. Since $u, d_2 \in X := H^1(\Omega)^d \cap L^\infty(\Omega)^d$ holds, there exists a constant R, which depends on u and d_2, such that $\|u+td_2\|_{L^\infty(\Omega)^d} \le R$ for $t \in [-1,1]$ and therefore also $\|u(x)+td_2(x)\| \le R$ for almost all $x \in \Omega$ and $t \in [-1,1]$. Hence, by (A.6) and using $\|\nabla u + t\nabla d_2\|^2 \le 2(\|\nabla u\|^2 + \|\nabla d_2\|^2)$, we obtain the following estimate, which

holds almost everywhere on Ω:

$$\begin{aligned}|g(x,u,d_1,d_2,t)| \le C(R)\Big[&(1+2(\|\nabla u(x)\|^2+\|\nabla d_2(x)\|^2))\|d_1\|_{L^\infty(\Omega)^d}\|d_2\|_{L^\infty(\Omega)^d}\\ &+\sqrt{1+2\|\nabla u(x)\|^2+2\|\nabla d_2(x)\|^2}\Big(\|\nabla d_2(x)\|\|d_1\|_{L^\infty(\Omega)^d}+\|\nabla d_1(x)\|\|d_2\|_{L^\infty(\Omega)^d}\Big)\\ &+\|\nabla d_1(x)\|\|\nabla d_2(x)\|\Big].\end{aligned}$$

The right-hand side is integrable, which follows from $\|\nabla u\|, \|\nabla d_2\|, \|\nabla d_1\| \in L^2(\Omega)$. This justifies our formal argument, and $J''(u)[d_1,d_2] = \varphi'(0)$ holds. Using Hölder's inequality, the boundedness of the differential $J''(u)$ follows easily (with a different $C(R)$)

$$\sup_{\|d_1\|_X=1,\|d_2\|_X=1} |J''(u)[d_1,d_2]| \le C(R)(\mu + \|\nabla u\|_{L^2(\Omega)^{dN}}).$$

Here, $\mu > 0$ is a constant which does not depend on d_1, d_2 and u. As in the other case, this shows that J is twice Gâteaux differentiable.

It is left to show that $u \mapsto J''(u)[d,d]$ is continuous. Again, we first consider that the assumptions (A.5) are satisfied. We show exemplary that the function

$$G_d(u) := \int_\Omega f_{uz}(x,u,\nabla u)[d,\nabla d]\,\mathrm{d}x = \int_\Omega d^T \nabla^2_{uz} f(x,u,\nabla u)\nabla d\,\mathrm{d}x$$

is continuous. The continuity of the other parts of $J''(u)[d,d]$ can be shown in the same way. Note that we use the notation $\nabla^2_{uz} f$ to refer to the matrix representation of f_{uz}. Obviously, the function $\phi(x,(u,\nabla u)) := \nabla^2_{uz} f(x,u,\nabla u)$ is a bounded Carathéodory function. Hence, we can apply Lemma A.7 which yields the continuity of G_d.

Now assume that (A.6) are satisfied instead. As in the other case, one uses Lemma A.7 to prove the continuity of

$$u \mapsto \int_\Omega f_{zz}(x,u,\nabla u)[\nabla d,\nabla d]\,\mathrm{d}x.$$

It is left to show the continuity of the function

$$H(u) := \int_\Omega \Big(f_{uu}(x,u,\nabla u)[d,d] + 2f_{uz}(x,u,\nabla u)[d,\nabla d]\Big)\,\mathrm{d}x.$$

For this, let (u_k) be a sequence that converges to u strongly in $H^1(\Omega)^N$. Using Hölder's inequality, we estimate

$$\begin{aligned}|H(u_k)-H(u)| \le \int_\Omega \Big(&\|\nabla^2_{uu} f(x,u,\nabla u) - \nabla^2_{uu} f(x,u_k,\nabla u_k)\|\|d(x)\|^2\\ &+2\|\nabla^2_{uz} f(x,u,\nabla u) - \nabla^2_{uz} f(x,u_k,\nabla u_k)\|\|d\|\|\nabla d\|\Big)\,\mathrm{d}x\\ \le C\Big(&\|d\|^2_{L^\infty(\Omega)}\int_\Omega \|\nabla^2_{uu} f(x,u,\nabla u) - \nabla^2_{uu} f(x,u_k,\nabla u_k)\|\,\mathrm{d}x\\ &+2\|d\|_{L^\infty(\Omega)}\|\nabla d\|_{L^2(\Omega)}\Big(\int_\Omega \|\nabla^2_{uz} f(x,u,\nabla u) - \nabla^2_{uz} f(x,u_k,\nabla u_k)\|^2\,\mathrm{d}x\Big)^{1/2}\Big).\end{aligned}$$

From Lemma A.6, we obtain that the operators

$$\Phi_1 \colon L^2(\Omega)^{N+dN} \to L^1(\Omega)^{N\times N},\ \Phi_1((u,\nabla u))(x) := \nabla^2_{uu} f(x,u(x),\nabla u(x))$$

and

$$\Phi_2 \colon L^2(\Omega)^{N+dN} \to L^2(\Omega)^{N\times dN},\ \Phi_2((u,\nabla u))(x) := \nabla^2_{uz} f(x,u(x),\nabla u(x))$$

are continuous. As a consequence, we have $|H(u_k) - H(u)| \to 0$ as $u_k \to u$. This finishes the proof. □

Remark A.3 Similar to Remark A.2, we can use the Sobolev embeddings to weaken the assumptions on the second derivatives. In the case $d = 2$, we can substitute (A.5) by

$$\begin{aligned}
&\|f_{uu}(x,u,z)\| \le C(g_1(x) + \|u\|^{q-2} + \|z\|^{2-4/q}), && g_1 \in L^{q/(q-2)}(\Omega)\\
&\|f_{uz}(x,u,z)\| \le C(g_2(x) + \|u\|^{(q-2)/2} + \|z\|^{1-2/q}), && g_2 \in L^{2q/(q-2)}(\Omega)\\
&\|f_{zz}(x,u,z)\| \le C,
\end{aligned}$$

with $2 < q < \infty$.

A.5. Existence of optimal points

We summarize some results which we use to discuss the existence of solutions of infinite dimensional optimization problems.

Definition A.3 Let U be a normed vector space. A function $f \colon U \to \mathbb{R}$ is called *coercive* in a set $C \subset U$ iff

$$\lim_{\substack{\|u\|_U \to \infty \\ u \in C}} f(u) = \infty.$$

Definition A.4 Let U be a normed vector space. A function $f \colon U \to \mathbb{R}$ is called *weakly lower semicontinuous* iff

$$u_k \rightharpoonup u \Rightarrow f(u) \le \lim_{k\to\infty} \inf f(u_k).$$

Theorem A.6 *Let U be a reflexive Banach space and f weakly lower semicontinuous. Furthermore, let $C \subset U$ be a nonempty and weakly sequentially closed subset. If C is bounded or f is coercive in C, then f takes its minimum in C.*

Proof Since C is nonempty, the infimum of f in C exists and hence we find a minimizing sequence $(u_k) \subset C$. We have assumed that C is bounded or f is coercive and due to this (u_k) must be bounded. Due to Theorem A.3, every bounded sequence contains a weakly convergent subsequence. Therefore, there exists (u_{k_j}) with $u_{k_j} \rightharpoonup u^*$ as $j \to \infty$. The feasible set C is assumed to be weakly closed and therefore $u^* \in C$. Finally, from the fact that F is weakly lower semicontinuous follows $f(u^*) \le \lim_{j\to\infty} \inf f(u_{k_j})$ and hence $f(u^*) \le f(u_k)$. Therefore, $u^* \in C$ is the minimum. □

Corollary A.1 *Let U be a reflexive Banach space, $C \subset U$ be a closed and convex set and f be continuous, convex and coercive in C. Then f takes its minimum in C.*

Proof By Lemma A.2, it follows that C is weakly sequentially closed. Furthermore, one can show that under the assumptions f is weakly lower semicontinuous (see for instance [Wer07, Lemma III.5.9]). Hence, the assertion follows directly from the preceding lemma. □

A.5.1. Weakly lower semicontinuity of variational integrals

Theorem A.7 *Let $\Omega \subset \mathbb{R}^d$ be an open and bounded set with Lipschitz-continuous boundary. Furthermore, let $f(x,u,z)\colon \Omega \times \mathbb{R} \times \mathbb{R}^d$ be such that*

1. *$f(x,u,z)$ is a* Carathéodory *function, i.e., f is measurable in x for every (u,z) and continuous in (u,z) for almost every $x \in \Omega$.*
2. *$z \mapsto f(x,u,z)$ is convex for almost every $x \in \Omega$ and for every $u \in \mathbb{R}$.*
3. *f is bounded below.*

Then the function

$$J(u) := \int_\Omega f(x,u,\nabla u)\,\mathrm{d}x$$

is weakly lower semicontinuous in $W^{1,1}(\Omega)$.

Proof See for instance [Giu03, Corollary 4.1]. □

Remark A.4 In the preceding theorem, the condition f is bounded below can be weakened by demanding that instead

$$f(x,u,z) \geq -C(|z|^m + |u|^k + g(x))$$

with $C > 0$, $g \in L^1(\Omega)$ if $u \in L^k(\Omega)$, $k \geq 1$, and $\nabla u \in L^p(\Omega)$, $m < p$, holds.

A.5.2. Regularity

Theorem A.8 *Let $\Omega \subset \mathbb{R}^d$ be an open and bounded set with Lipschitz-continuous boundary. Furthermore, let $f(x,u,z)\colon \Omega \times \mathbb{R} \times \mathbb{R}^d$ be a Carathéodory function that satisfies the growth condition*

$$\varepsilon|z|^p - b(x)|u|^k - a(x) \leq f(x,u,z) \leq L|z|^p + b(x)|u|^k + a(x)$$

where $\varepsilon, L > 0$, $1 < p \leq k < p^* := \frac{pd}{d-p}$ and a, b are non-negative functions belonging to $L^s(\Omega)$, $s > n/p$, and $L^t(\Omega)$, $t > \frac{p^*}{p^*-k}$. Assume that γ is Hölder continuous on $\partial\Omega$ and $\phi \in W^{1,p}(\Omega)$ with $tr(\phi) \leq \gamma$ a.e.. Then each solution u^* of the problem

$$\min_{u \in \mathcal{C}} \int_\Omega f(x,u,\nabla u)\,\mathrm{d}x$$

where $\mathcal{C} = \{u \in W^{1,p}(\Omega) \,|\, tr(u) = \gamma,\, u \geq \phi\}$ is Hölder continuous in $\overline{\Omega}$.

Proof Follows from Example 6.4 and Theorem 7.8 in [Giu03]. □

Acknowledgment

Writing this thesis would not have been possible without the help and the contribution of a lot of people. It is a great pleasure for me to thank all of them.

First and foremost, I would like to express my gratitude to my supervisor Prof. Dr. Michael Ulbrich for introducing me to the subject of nonlinear optimization, for making it possible for me to write this thesis and for his support over the last years.

I started this thesis at the University of Hamburg and I wish to thank my former colleagues, in particular Matthias Kabel, for the good time and the interesting discussions we had.

Furthermore, I would like to thank all the members of our working group for the excellent and kind atmosphere and the daily support that I have received. It was a pleasure to work with all of you. In particular, I thank Florian Kruse, with whom I shared my room, for answering countless questions and the fruitful and fun time we had together in 03.08.061.

I am indebted to all the proofreaders, in particular Benjamin Crost, who had the difficult task to bring my English spelling and style in due form.

I would also like to thank my mother Gilla and my sister Mariella for their never-ending help and ongoing support.

Very special thanks go to my dear friend Oliver for his friendship that is so important for me.

Last but by no means least, I thank my girlfriend Jessica for her loving support and willingness to support me the whole time in Munich. I love you.

Bibliography

[ACM91] Brett M. Averick, Richard G. Carter, and Jorge J. Moré, *The MINPACK-2 test problem collection*, Technical Memorandum ANL/MCS-TM-150, Argonne National Laboratory, Argonne, IL, USA, 1991.

[AL01] Natalia M. Alexandrov and Robert Michael Lewis, *An overview of first-order model management for engineering optimization*, Optim. Eng. **2** (2001), no. 4, 413–430.

[All86] Allgower, E. L. and Böhmer, K. and Potra, F. A. and Rheinboldt, W. C., *A mesh-independence principle for operator equations and their discretizations*, SIAM J. Numer. Anal. **23** (1986), 160–169.

[Alt06] Hans Wilhelm Alt, *Lineare Funktionalanalysis. Eine anwendungsorientierte Einführung. 5th revised ed*, Springer, Berlin, 2006.

[AN95] Ya.I. Alber and A.I. Notik, *On some estimates for projection operators in Banach spaces*, Comm. Appl. Nonlinear Anal. **2** (1995), no. 1, 47–55.

[Bad06] Lori Badea, *Convergence rate of a Schwarz multilevel method for the constrained minimization of nonquadratic functionals*, SIAM J. Numer. Anal. **44** (2006), no. 2, 449–477.

[BC83] Achi Brandt and Colin W. Cryer, *Multigrid algorithms for the solution of linear complementarity problems arising from free boundary problems*, SIAM J. Sci. Stat. Comput. **4** (1983), 655–684.

[BHT09] Michele Benzi, Eldad Haber, and Lauren Taralli, *Multilevel algorithms for large-scale interior point methods*, SIAM J. Sci. Comput. **31** (2009), no. 6, 4152–4175.

[Bor05] Alfio Borzì, *On the convergence of the MG/OPT method*, Proc. Appl. Math. Mech. **5** (2005), no. 1, 735–736.

[BPV00] James H. Bramble, Joseph E. Pasciak, and Panayot S. Vassilevski, *Computational scales of Sobolev norms with application to preconditioning*, Math. Comp. **69** (2000), no. 230, 463–480.

[BPX90] James H. Bramble, Joseph E. Pasciak, and Jinchao Xu, *Parallel multilevel preconditioners*, Math. Comp. **55** (1990), no. 191, 1–22.

[BR82] Randolph E. Bank and Donald J. Rose, *Analysis of a multilevel iterative method for nonlinear finite element equations*, Math. Comp. **39** (1982), 453–465.

[Bra77] Achi Brandt, *Multi-level adaptive solutions to boundary-value problems*, Math. Comp. **31** (1977), 333–390.

[Bra95] James H. Bramble, *On the development of multigrid methods and their analysis*, Mathematics of Computation 1943-1993: A Half-Century of Computational Mathematics (Walter Gautschi, ed.), Proc. Symp. Appl. Math., vol. 48, American Mathematical Society, 1995, pp. 5–19.

[Bra07] Dietrich Braess, *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*, 3rd ed., Cambridge University Press, 2007.

[BS08] Susanne C. Brenner and Ridgway L. Scott, *The mathematical theory of finite element methods*, 3rd ed., Texts in Applied Mathematics, vol. 15, Springer, 2008.

[BS09] Alfio Borzì and Volker Schulz, *Multigrid methods for PDE optimization*, SIAM Rev. **51** (2009), no. 2, 361–395.

[BTW03] Lori Badea, Xue-Cheng Tai, and Junping Wang, *Convergence rate analysis of a multiplicative Schwarz method for variational inequalities*, SIAM J. Numer. Anal. **41** (2003), no. 3, 1052–1073.

[BVW03] Peter N. Brown, Panayot S. Vassilevski, and Carol S. Woodward, *On mesh-independent convergence of an inexact Newton–multigrid algorithm*, SIAM J. Sci. Comput. **25** (2003), no. 2, 570–590.

[BX91] James H. Bramble and Jinchao Xu, *Some estimates for a weighted L^2 projection*, Math. Comp. **56** (1991), no. 194, 463–476.

[BY93] Folkmar Bornemann and Harry Yserentant, *A basic norm equivalence for the theory of multilevel methods*, Numer. Math. **64** (1993), no. 4, 455–476.

[BZ00] James H. Bramble and Xuejun Zhang, *The analysis of multigrid methods*, Handbook of numerical analysis (P. G. Ciarlet and J. L. Lions, eds.), vol. VII, North-Holland / Elsevier, 2000, pp. 173–415.

[Car99] Carsten Carstensen, *Quasi-interpolation and a posteriori error analysis in finite element methods*, Math. Model. Numer. Anal. **33** (1999), no. 6, 1187–1202.

[CG82] Philippe G. Ciarlet and G. Geymonat, *Sur les lois de comportement en élasticité non-linéaire compressible*, C. R. Math. Acad. Sci. Paris **295** (1982), no. II, 423–426.

[CGT00] Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint, *Trust-region methods*, SIAM, 2000.

[CHGK93] Kwan J. Chang, Raphael T. Haftka, Gary L. Giles, and Pi-Jen Kao, *Sensitivity-based scaling for approximating structural response*, J. Aircraft **30** (1993), 283–288.

[Cia78] Philippe G. Ciarlet, *The finite element method for elliptic problems*, Studies in mathematics and its applications, vol. 4, North-Holland, 1978.

[Cia88] ———, *Mathematical elasticity. Volume I: Three-dimensional elasticity*, Studies in mathematics and its applications, vol. 20, North-Holland, 1988.

[CL96] Thomas F. Coleman and Yuying Li, *An interior trust region approach for nonlinear minimization subject to bounds*, SIAM J. Optim. **6** (1996), no. 2, 418–445.

[DDW75] Jim jun. Douglas, Todd Dupont, and Lars Wahlbin, *The stability in L^q of the L^2-projection into finite element function spaces*, Numer. Math. **23** (1975), 193–197.

[DMM04] Elizabeth D. Dolan, Jorge J. Moré, and Todd S. Munson, *Benchmarking optimization software with COPS 3.0*, Tech. Report ANL/MCS-TM-273, Argonne National Laboratory, Feb. 2004.

[DMS00] Thomas Dreyer, Bernd Maar, and Volker Schulz, *Multigrid optimization in applications*, J. Comput. Appl. Math. **120** (2000), no. 1-2, 67–84.

[DW97] Peter Deuflhard and Martin Weiser, *Local inexact newton multilevel FEM for nonlinear elliptic problems*, Computational science for the 21st century (M.-O. Bristeau et al., ed.), John Wiley & Sons., 1997, pp. 129–138.

[Fed61] Radii P. Fedorenko, *A relaxation method for solving elliptic difference equations*, U.S.S.R. Comput. Math. Math. Phys. **1** (1961), no. 4, 1092–1096.

[Fed64] ———, *The speed of convergence of one iterative process*, U.S.S.R. Comput. Math. Math. Phys. **4** (1964), no. 3, 1092–1096.

[Giu03] Enrico Giusti, *Direct methods in the calculus of variations*, World Scientific, 2003.

[GK09a] Christian Gross and Rolf Krause, *On the convergence of recursive trust-region methods for multiscale nonlinear optimization and applications to nonlinear mechanics*, SIAM J. Numer. Anal. **47** (2009), no. 4, 3044–3069.

[GK09b] Carsten Gräser and Ralf Kornhuber, *Multigrid methods for obstacle problems*, J. Comput. Math. **27** (2009), no. 1, 1–44.

[Glo84] Roland Glowinski, *Numerical methods for nonlinear variational problems*, Springer Series in Computational Physics, Springer, 1984.

[GMS+10] Serge Gratton, Mélodie Mouffe, Annick Sartenaer, Philippe L. Toint, and Dimitri Tomanos, *Numerical experience with a recursive trust-region method for multilevel nonlinear bound-constrained optimization*, Optim. Methods Softw. **25** (2010), no. 3, 359–386.

[GMTWM08] Serge Gratton, Mélodie Mouffe, Philippe L. Toint, and Melissa Weber-Mendonça, *A recursive ℓ_∞-trust-region method for bound-constrained nonlinear optimization*, IMA J. Numer. Anal. **28** (2008), no. 4, 827–861.

[GST08] Serge Gratton, Annick Sartenaer, and Philippe L. Toint, *Recursive trust-region methods for multiscale nonlinear optimization*, SIAM J. Optim. **19** (2008), no. 1, 414–444.

[GVL96] Gene Golub and Charles F. Van Loan, *Matrix computations*, 3rd ed., The Johns Hopkins Univ. Press, 1996.

[Hac85] Wolfgang Hackbusch, *Multi-grid methods and applications*, Springer Series in Computational Mathematics, vol. 4, Springer, 1985.

[Hac92] ———, *Elliptic differential equations: theory and numerical treatment*, Springer Series in Computational Mathematics, vol. 18, Springer, 1992.

[HL77] Ivan Hlavacek and Jan Lovisek, *A finite element analysis for the Signorini problem in plane elastostatics*, Appl. Math. **22** (1977), 215–228.

[Hop87] Ronald H. W. Hoppe, *Multigrid algorithms for variational inequalities*, SIAM J. Numer. Anal. **24** (1987), 1046–1065.

[HPUU09] Michael Hinze, Rene Pinnau, Michael Ulbrich, and Stefan Ulbrich, *Optimization with PDE constraints*, Springer, 2009.

[HR89] Wolfgang Hackbusch and Arnold Reusken, *Analysis of a damped nonlinear multilevel method*, Numer. Math. **55** (1989), no. 2, 225–246.

[HW05] Stefan Hüeber and Barbara Wohlmuth, *A primal-dual active set strategy for nonlinear multibody contact problems*, Comput. Methods Appl. Mech. Engrg. **194** (2005), no. 27-29, 3147–3166.

[IT79] Aleksandr D. Ioffe and Vladimir M. Tihomirov, *Theory of extremal problems*, Studies in Mathematics and its Applications, vol. 6, North-Holland, 1979.

[JOP$^+$] Eric Jones, Travis Oliphant, Pearu Peterson, et al., *SciPy: Open source scientific tools for Python*, 2001–.

[KK01] Ralf Kornhuber and Rolf Krause, *Adaptive multigrid methods for Signorini's problem in linear elasticity*, Comput. Vis. Sci. **4** (2001), no. 1, 9–20.

[KO88] Noboru Kikuchi and John T. Oden, *Contact problems in elasticity: A study of variational inequalities and finite element methods*, SIAM Studies in Applied Mathematics, vol. 8, SIAM, 1988.

[Kor94] Ralf Kornhuber, *Monotone multigrid methods for elliptic variational inequalities I*, Numer. Math. **69** (1994), no. 2, 167–184.

[Kor97] ———, *Adaptive monotone multigrid methods for nonlinear variational problems*, Advances in numerical mathematics, Teubner, 1997.

[Kor01] ———, *Nonlinear multigrid techniques*, Theory and numerics of differential equations (J. F. Blowey, J. P. Coleman, and A. W. Craig, eds.), Springer, 2001, pp. 179–229.

[LMW$^+$11] Anders Logg, Kent-Andre Mardal, Garth N. Wells, et al., *Automated solution of differential equations by the finite element method*, Lecture notes in computational science and engeneering, vol. 84, Springer, 2011.

[LN05] Robert Michael Lewis and Stephen G. Nash, *Model problems for the multigrid optimization of systems governed by differential equations*, SIAM J. Sci. Comput. **26** (2005), no. 6, 1811–1837.

[Man84] Jan Mandel, *A multilevel iterative method for symmetric, positive definite linear complementarity problems*, Appl. Math. Optim. **11** (1984), 77–95.

[Mor88] Jorge J. Moré, *Trust regions and projected gradients*, System modelling and optimization (M. Iri and K. Yajima, eds.), Lect. Notes Control Inf. Sci., no. 113, Springer-Verlag, 1988, pp. 1–13.

[Mor90] ———, *A collection of nonlinear model problems*, Computational Solution of Nonlinear Systems of Equations (Eugene L. Allgower and Kurt Georg, eds.), Lect. Appl. Math., vol. 26, 1990, pp. 723–762.

[Mor08] Charles B. Morrey Jr., *Multiple integrals in the calculus of variations*, Classics in Mathematics., Springer, 2008.

[Nas00] Stephen G. Nash, *A multigrid approach to discretized optimization problems*, Optim. Methods Softw. **14** (2000), no. 1-2, 99–116.

[Neu97] John W. Neuberger, *Sobolev gradients and differential equations*, Lecture notes in mathematics, vol. 1670, Springer, 1997.

[Nol93] Dominikus Noll, *Second order differentiability of integral functionals on Sobolev spaces and L^2-spaces*, J. Reine Angew. Math. **436** (1993), 1–17.

[Osw94] Peter Oswald, *Multilevel finite element approximation. Theory and applications*, Teubner Skripten zur Numerik, Teubner, 1994.

[PMX98] P. M. Pardalos, T. Mavridou, and J. Xue, *The graph coloring problem: A bibliographic survey*, vol. 2, pp. 331–395, Kluwer Academic Publishers, 1998.

[Pow70] Michael J. D. Powell, *A new algorithm for unconstrained optimization*, Nonlinear Programming (J. B. Rosen, O. L. Mangasarian, and K. Ritter, eds.), Publ. Math. Res. Center, Univ. Wisconsin–Madison, no. 25, Academic Press, 1970, pp. 31–65.

[Saa03] Yousef Saad, *Iterative methods for sparse linear systems*, 2nd ed., SIAM, 2003.

[Tai03] Xue-Cheng Tai, *Rate of convergence for some constraint decomposition methods for nonlinear variational inequalities*, Numer. Math. **93** (2003), no. 4, 755–786.

[Toi88] Philippe L. Toint, *Global convergence of a class of trust-region methods for nonconvex minimization in Hilbert space*, IMA J. Numer. Anal. **8** (1988), no. 2, 231–252.

[TOS01] Ulrich Trottenberg, Cornelis W. Oosterlee, and Anton Schüller, *Multigrid. With guest contributions by A. Brandt, P. Oswald, K. Stüben*, Elsevier Academic Press, 2001.

[Ulb01] Michael Ulbrich, *Nonmonotone trust-region methods for bound-constrained semismooth equations with applications to nonlinear mixed complementarity problems*, SIAM J. Optim. **11** (2001), no. 4, 889–917.

[UUH99] Michael Ulbrich, Stefan Ulbrich, and Matthias Heinkenschloss, *Global convergence of trust-region interior-point algorithms for infinite-dimensional nonconvex minimization subject to pointwise bounds*, SIAM J. Control Optim. **37** (1999), no. 3, 731–764.

[Var62] Richard S. Varga, *Matrix iterative analysis*, Prentice-Hall Series in Automatic Computation., Prentice-Hall, 1962.

[Wer07] Dirk Werner, *Funktionalanalysis*, 6th ed., Springer, 2007.

[Wes92] Peter Wesseling, *An introduction to multigrid methods*, A Wiley-Interscience Series of Texts, Monographs & Tracts., John Wiley & Sons Ltd., 1992.

[WG09] Zaiwen Wen and Donald Goldfarb, *A line search multigrid method for large-scale nonlinear optimization*, SIAM J. Optim. **20** (2009), no. 3, 1478–1503.

[WSD05] Martin Weiser, Anton Schiela, and Peter Deuflhard, *Asymptotic mesh independence of Newton's method revisited*, SIAM J. Numer. Anal. **42** (2005), no. 5, 1830–1845.

[Xu92] Jinchao Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev. **34** (1992), no. 4, 581–613.

[YD06] Irad Yavneh and Gregory Dardyk, *A multilevel nonlinear method*, SIAM J. Sci. Comput. **28** (2006), no. 1, 24–46.

[Yos80] Kosaku Yosida, *Functional analysis*, 6th ed., Grundlehren der mathematischen Wissenschaften, vol. 123, Springer, 1980.

[Yse93] Harry Yserentant, *Old and new convergence proofs for multigrid methods*, Acta Numer. **2** (1993), 285–326.

[Zei86] Eberhard Zeidler, *Nonlinear functional analysis and its applications. I: Fixed-point theorems*, Springer, 1986.

[Zei90] ———, *Nonlinear functional analysis and its applications. II/B: Nonlinear monotone operators*, Springer, 1990.

[Zha92] Xuejun Zhang, *Multilevel Schwarz methods*, Numer. Math. **63** (1992), no. 4, 521–539.