Inaugural – Dissertation
zur
Erlangung der Doktorwürde
der
Naturwissenschaftlich–Mathematischen Gesamtfakultät
der
Ruprecht–Karls–Universität
Heidelberg

vorgelegt von
Diplom–Physiker Wolfgang Bangerth
aus Ostfildern

Tag der mündlichen Prüfung: 12. Juli 2002

Adaptive Finite Element Methods for the Identification of Distributed Parameters in Partial Differential Equations

Gutachter: Prof. Dr. R. Rannacher
Prof. Dr. H. G. Bock

Contents

Introduction

When quantities cannot be measured directly, parameter estimation techniques come into play: with these, the unknown quantity is determined from measurements of observables. This work deals with problems where the relation between the observables and the desired information is a partial differential equation. Such parameter estimation problems are then commonly referred to as *Inverse Problems*.

Inverse problems have vast applications in science and engineering. In this work, we consider problems where internal properties of media are of interest, which, however, are often not accessible directly. For example, in some applications we are interested in determining the internal elastic composition of bodies without destroying it, or would like to know the underground structure in search of oil without actually drilling. These quantities appear as coefficients in the partial differential equations (henceforth abbreviated by *PDE*) which are used to describe the response of the media to forces, and the determination of these coefficient naturally leads to inverse problems.

From a numerical point of view, inverse problems involving partial differential equations are very challenging: unlike nonlinear partial differential equations, they do not only require the solution of *one* or few linearized subproblems in each nonlinear step, but *many*. Since we are looking for distributed parameters which may be discretized by thousands or tens of thousands degrees of freedom, the number of linearized subproblems in each nonlinear step may be several hundreds or thousands. As an example, the transmission tomography application discussed in Section 5.6 required a total of 2008 CG iterations, accumulated over some 80 Newton steps. Since 32 experiments were used, this means a total of roughly 130,000 solutions of a Helmholtz equation. Computational considerations are therefore of outstanding importance in the design of algorithms to solve such problems.

Consequently, the goal of this work is the development of techniques for the *efficient* numerical solution of such inverse problems, based on adaptive finite element methods. After the statement of the problem in Chapter 1, we will derive *a posteriori error estimates* for inverse problems in Chapter 2, both for natural "energy type" quantities as well as for general functionals, and demonstrate their efficiency. Although adaptivity and error estimation are now commonly accepted in the numerical solution of partial differential equations, they have not yet found their way into the solution of inverse problems. These techniques are thus new to this field and promise a significant gain in efficiency compared to present state-of-the-art algorithms.

A second, new aspect of this work is the inclusion of bounds into the solution process in Chapter 3. In practical applications, physical upper and lower bounds on possible values of the unknown coefficients are usually available, either from prior knowledge of the particular case under investigation, or from extremal material properties existing in nature. For example, when identifying the underground structure from seismic measurements, densities of rocks will be between approximately 1 g/cm^3 (water) and 22 g/cm^3 (osmium and alike metals). In practice, such bounds are usually much tighter, and alike bounds are available for other properties as well, such as elasticity coefficients. The efficient inclusion of such bounds is discussed in Chapter 3 where we develop an *Active Set Method* in a continuous setting and show its efficiency in enhancing stability of identified coefficients.

In Chapter 4, we extend the problems under consideration to the case that more than just one measurement is available. This can be favorably used to suppress the effects of measurement noise, and examples of this are shown. It also allows to solve certain classes of problems in which one measurement is not sufficient to identify the unknown coefficient. Beyond the already high computational requirements for distributed parameter identification in PDEs, multiple measurements increase it even more. This requires using specialized algorithms tailored to the problem. However, their structure allows for efficient parallelization strategies, for example using clusters of computers. The work required for each of the subproblems associated with one measurement is thus distributed to different computers. The structure of a program doing this will be introduced in Chapter 4.

The techniques developed thus far at the Laplace equation will be applied to parameter identification problems for the Helmholtz equation in Chapter 5. Since Helmholtz's equation is the frequency domain version of the wave equation, parameter estimation for this type of problems has many applications in geophysics. It will be shown that adaptive techniques and error estimation work in this context as well, and that they lead to very efficient schemes. The most complex problems of this work will be considered in this chapter.

We conclude with an outlook on the challenges of inverse problems that are not, or only briefly, touched in this work.

Two prototypical applications

The techniques developed in this thesis should be considered in view of actual applications. To this aim, we introduce two prototypical applications. The first one, nondestructive testing, tries to determine the elastic properties of a material by subjecting it to a known force, and measuring the resulting deflection. The second, electrical impedance tomography, uses electrical potentials applied to the boundary of a body to image its interior.

Nondestructive testing. Assume we want to infer the stiffness properties of a body without taking it apart or destroying it otherwise, for example because it is precious or because an assessment of the body is required before it is

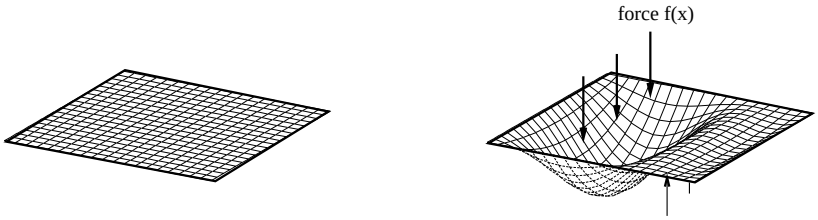


Figure 1: *Principle of nondestructive testing by application of forces. Left: Membrane in rest state. Right: Membrane deflected in reaction to an applied external force field* $f(\mathbf{x})$.

deployed to use. This frequently occurs in quality control of parts in aerospace industries, and many other applications.

The idea of the method applied to a membrane of spatially varying elastic properties is then as follows (see Figure 1): knowing the rest state of the membrane in the absence of external forces, we want to infer the desired material properties by measuring the deflection after applying a force of known spatial distribution and strength.

A mathematically concise definition of this problem will be given in Chapter 1, so we only present a sketch of a formulation: For the membrane under consideration, assume that its deflection u is described by a Poisson equation

$$-\nabla \cdot (a \nabla u) = f,$$

where f is the applied body force and $a = a(\mathbf{x})$ the spatially varying coefficient we would like to recover. For a complete model, the equation is of course augmented by suitable boundary conditions.

While we do not know the coefficient, we have measured the deflection u of the membrane under action of the applied force. We denote this measurement by z. Since we can compute a deflection u for each possible coefficient (bounded away from zero), the problem of parameter identification can then be stated as follows: *find that coefficient for which the corresponding deflection* u *matches the measured deflection* z *best.* Methods for finding this coefficient will be discussed in the next chapter.

Electrical impedance tomography. Another, closely related problem is the determination of the electrical properties of a body from measurements at its boundary. This has applications in the detection of interior cracks in metallic parts in aerospace industries, but is also envisaged as an imaging technique in medical applications. Here, see Figure 2, one tries to infer the internal electrical conductivities of a body by applying electrostatic potentials to its boundary; the observable quantity is then the resulting electric field at the boundary, which depends on the potentials and the internal composition of the body. From this one hopes to invert for the interior. Because this method tries to *see* into the

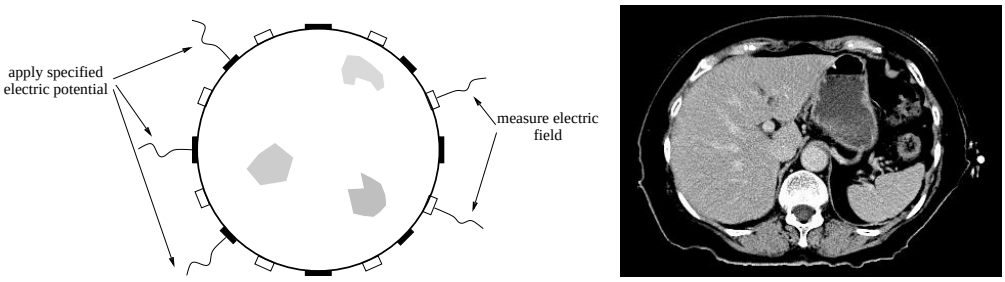


Figure 2: *Principle of electrical impedance tomography: subject a body to specified electrical potentials at its boundary, and measure the resulting electrical fields. Left: Scheme of measurements. Right: computer tomographic image of the human upper body, for which electrical impedance tomography could be an alternative imaging technique.*

body only from measurements outside of it, it is often called *electrical impedance tomography.*

Mathematically speaking, we now have a Laplace equation describing the electric potential with a variable coefficient which we would like to recover. Instead of body forces, we now have Dirichlet boundary values (i.e. the applied surface potential) as sources, and the Neumann boundary values (i.e. the electrical field at the surface) as observables. In this thesis, we do not discuss this particular problem for the Laplace problem, but for the Helmholtz equation.

Problems related to this one occur in a large number of applications. It is a recurring theme in geophysics (see the books by Tarantola [63] and Parker [54]), where, for example, measurements of the gravimetric potential are used to obtain information about underground structures associated with mass distribution anomalies. If we extend the problems to time dependent ones, the seismic inversion problem is also of this type: there the goal is to obtain information about the underground from the measurement of seismic signals. Important applications of this are earthquake prediction and oil reservoir identification.

What is the solution of an inverse problem?

In this work, we try to identify the *maximum likelihood point* of a problem. To keep with the membrane example above, this means that we seek the single one coefficient for which the predicted deflection matches the measured one best. However, this is in some sense a rather restricted point of view: since the measurement usually contains noise, any other noise realization of the measurement would be equally valid, and for each we might get a different "best" coefficient.

The most appropriate definition of a solution therefore would be a probability distribution in coefficient space: for each noisy measurement occurring with a certain probability, assign this probability to the corresponding "best" coefficient.

For most parameter identification problems involving partial differential equations, recovering this probability density exceeds today's computational possibilities by far. We therefore restrict our point of view to the identification

of *one* distributed coefficient function, and note that this is also appropriate for the case of small noise, since then the probability density is approximately Gaussian with peak at this one coefficient and computable width. This restriction must, however, be kept in mind when thinking about inverse problems. For further discussions in this direction, see the outlook section of this work (page 105), and in particular the book by Tarantola [63].

A word on notation

The scientific communities concerned with the numerical solution of partial differential equations, and with optimization maintain different, incompatible conventions of notation. For example the state variable is commonly named u in numerical analysis, while it is denoted by x, or $y(x)$, in optimization theory. Since this work is mainly concerned with numerical aspects, in particular the finite element approximation of optimization problems, we will use the notation common in numerical analysis.

Acknowledgments

I feel deeply obliged to the people who accompanied me here, and without whom this work could not have been undertaken. In particular, I want to thank my parents and Heike for their support, and Rainer and Guido for their friendship; Professor Rolf Rannacher for his support and giving me the freedom to choose my subjects of research; Ralf and the rest of the institute for the atmosphere here. This work was funded by the Graduiertenkolleg "Modellierung und Wissenschaftliches Rechnen in Mathematik und Naturwissenschaften" and the Sonderforschungsbereich 359 "Reaktive Strömungen, Diffusion und Transport".

Chapter 1

Parameter estimation for elliptic problems

In this first chapter, we will give an outline of the way by which we intend to attack the problem of estimation of distributed parameters in elliptic partial differential equations. We first discuss the formal setting of the problem in mathematical terms, then formulate it as a constrained minimization problem for which we seek the stationary point of a Lagrangian.

This constrained problem is stated in a continuous setting in function spaces. For its solution, we employ Newton's method, again on a continuous level. The individual Newton steps are then discretized using a finite element method that differs from the approaches used in the available literature in that we use different meshes and shape functions for the different types of variables present.

The rest of the chapter is devoted to the discussion of the solution of the linear subproblems and theoretical questions regarding the framework outlined so far. The chapter closes with the definition of some benchmarks that will be used in later chapters.

As already mentioned in the introduction, the solutions we are seeking in this work – by requiring the stationarity of a Lagrangian – are maximum likelihood points in the model space. What we call *solution* to the inverse problem is thus only a certain aspect of it. We do not consider the identification of the full posterior probability density function in the model space which would require us to use significantly different techniques than we intend to discuss in this work, as for example Monte Carlo sampling. Questions like resolution and significance, or variances and cross-variances are therefore not covered and are left for future research. For more details about these questions, we refer to the book by Tarantola [63].

1.1 A model problem

This work is devoted to the identification of distributed coefficients in partial differential equation equations. A model diffusion problem involving the Laplace equation, as well as the necessary notation to describe it, is introduced in this section. This model problem will be used in all following chapters except

for the last one where identification problems for the Helmholtz equation are considered.

The problems considered here are of the following form: assume we have measurements z of certain physically observable quantities, such as displacements of a membrane, electrical fields at the surface of a body, or seismic signals. We know that these signals are caused by some sources f and g located in the interior and on the boundary of the domain, respectively, and that the physical system can be described by a partial differential equation that allows a unique solution u denoting the state the system is in. This equation depends on certain material properties of the system, denoted by the variable a, which cannot be observed directly, but which we would like to infer from the measurements. The task is then to find such model parameters a for which the output (i.e. the state u of the system or certain aspects of it) matches the observations best. We particularly assume that we are looking for spatially varying parameters $a = a(\mathbf{x})$.

In practical applications we often have additional knowledge. For example, information about the parameter of the form $a_0 \leq a \leq a_1$ may be available; these bounds occur since for model parameters such as elasticity coefficients, density, or attenuation, lower and upper bounds are readily constructible by considering the extreme cases for the materials of which the medium is composed. This information will be incorporated into the methods developed in this work if possible.

Given the above, a formulation of the problem in words may be as follows:

Problem 1.1. *Minimize the difference between u and z with respect to a given misfit functional by varying the parameters $a(\mathbf{x})$, under the constraint that at the solution $\{u^*, a^*\}$ the state equation is satisfied, and that $a_0 \leq a^* \leq a_1$.*

Below, one mathematical formulation of this problem will be stated, see Problem 1.7, and the resulting equations determining the solutions u^* and a^* are derived along with methods to solve them. We will frequently drop the asterisk at the solution if no confusion is possible.

While we use only one formulation of the parameter identification problem, we note that there are many which we do not touch here. Some of these are mentioned at the end of this section.

In order to state the problem of parameter identification in a concise way, we first define some notation for later use:

Definition 1.2 (Function spaces). *Denote by $H^p(\Omega)$ the usual Sobolev space of functions over the domain Ω which are in $L^2(\Omega)$ and have derivatives up to order p in $L^2(\Omega)$, see Yosida [68]. Let $H_0^1 = \{v \in H^1(\Omega) : v|_{\partial\Omega} = 0\}$, and define by $H^{-1}(\Omega) = H_0^1(\Omega)'$ its dual.*

Based on these spaces, let $H^{1/2}(\Gamma)$ denote the normed space of traces of $H^1(\Omega)$ functions on Γ, with norm induced by the trace operator (see Schwab [59]). Finally, let $\Gamma_D \subset \partial\Omega$ and define

$$V_g = \{v \in H^1(\Omega) : v|_{\Gamma_D} = g\},$$
$$V_0 = \{v \in H^1(\Omega) : v|_{\Gamma_D} = 0\},$$

with $g \in H^{1/2}(\Gamma_D)$.

Definition 1.3 (State equation). *Let* Ω *be a bounded, open subset of* $\mathbb{R}^n$ *and let*

$$\begin{aligned} -\nabla \cdot (a\nabla u) &= f, && \text{in } \Omega, \\ u &= g, && \text{on } \Gamma_D \subset \partial\Omega, \\ a\partial_n u &= 0, && \text{on } \Gamma_N = \partial\Omega - \Gamma_D \end{aligned}$$

be the elliptic differential equation for which we want to find the parameter a *from measurements* z *of the solution* u *. For simplicity, we assume* Ω *to be polygonal. The state equation is understood to be in the weak sense, i.e. we require that for* $u \in V_g$ *satisfies*

$$(a\nabla u, \nabla\varphi) - (f, \varphi) = 0 \qquad \forall \varphi \in V_0(\Omega), \tag{1.1}$$

where

$$\begin{aligned} &a \in \mathcal{A} = \{a \in L^\infty(\Omega) : 0 < \alpha \le a\}, \\ &f \in H^{-1}(\Omega), \\ &g \in H^{1/2}(\Gamma_D). \end{aligned}$$

In many applications, we will also be able to exploit physical knowledge about the parameter a. While for well-posedness of the state equation we only need that a is bounded away from zero, known material properties of the parameters often allow us to bound $a_0 \le a \le a_1$, with a_0, a_1 being constant or varying in space. We will include these bounds into the definition of $\mathcal{A}$:

Definition 1.4 (Parameter space). *Let the admissible set for the parameter be*

$$\mathcal{A} = \{a \in L^\infty(\Omega) : 0 < \alpha \le a_0(\mathbf{x}) \le a(\mathbf{x}) \le a_1(\mathbf{x}) < \infty\}.$$

Furthermore, we define the tangent cone to $\mathcal{A}$ *at position* a *by*

$$\begin{aligned} \mathcal{A}'[a] = \big\{\chi \in L^\infty :\ & \chi(\mathbf{x}) \ge 0 \text{ for } \mathbf{x} \in \{\mathbf{x} : a(\mathbf{x}) = a_0\}, \\ & \chi(\mathbf{x}) \le 0 \text{ for } \mathbf{x} \in \{\mathbf{x} : a(\mathbf{x}) = a_1\}\big\}. \end{aligned}$$

The problem we are concerned with in this work involves the minimization of the difference between the solution of an equation u and a measurement z. We will now define how we measure this difference:

Definition 1.5 (Misfit functionals). *Let* $u \in V_g$ *be the solution of* (1.1)*, and* $z \in \mathcal{M}$ *be the measurement. Let* $M : V_g \to \mathcal{M}$ *be a mapping from the space of solutions into the space of measurements* $\mathcal{M}$*. We will then measure the misfit between solution and measurement,*

$$m(Mu - z),$$

with a convex and continuous functional $m : \mathcal{M} \to \mathbb{R}_0^+$ *, normalized to* $m(0) = 0$.

We will frequently write $m(u-z)$ instead of $m(Mu-z)$ if M is simply the embedding of V_g into another space (e.g. into $\mathcal{M} = L^2(\Omega)$), or a canonical restriction (e.g. the restriction to a part $\Omega' \subset \Omega$, or the trace mapping from $H^1(\Omega)$ into $L^2(\Gamma)$ with some curve Γ).

The first and second derivative of $m(\cdot)$ at position $u-z$ will be denoted by $m'(u-z;\cdot)$ and $m''(u-z;\cdot,\cdot)$, respectively. If m is quadratic in its argument, $m''(u-z;\cdot,\cdot)$ does not depend on $u-z$.

Examples for misfit functionals corresponding to domain measurements are

$$m(u-z) = \tfrac{1}{2}\|u-z\|^2_{L^2(\Omega)}, \qquad \text{or} \qquad m(u-z) = \tfrac{1}{2}\|\nabla u - z\|^2_{L^2(\Omega)}.$$

These are used if measurements of the state variable or its gradient are available everywhere. Measurements on the boundary are also possible, as well as weighted norms. More complicated measurement functionals may be tailored to the statistical properties of measurement noise. Examples include L^1 norms of value or gradient, or smoothed variants thereof, such as Huber's or Ekblom's measure (see, for example, Amundsen [2] and Farquharson and Oldenburg [35]).

Due to noise in the measurement z we usually need to add a regularization term to the functional we want to minimize. Its form is stated in the following definition:

Definition 1.6 (Regularization functionals). *The regularization functionals used in this work are denoted by $r : \mathcal{A} \to \mathbb{R}_0^+$. They are assumed to be convex and differentiable, and normalized to $r(0) = 0$.*

Again, first and second derivatives are denoted by $r'(a;\cdot)$ and $r''(a;\cdot,\cdot)$, respectively. Common choices for $r(\cdot)$ include

$$r(a) = \tfrac{1}{2}\|a\|^2_{L^2(\Omega)}, \qquad \text{or} \qquad r(a) = \tfrac{1}{2}\|\nabla a\|^2_{L^2(\Omega)},$$

or again other functionals such as the ones mentioned above. In general, the choice of the regularization functional should be guided by physical insight into the problem at hand, as regularization should penalize certain undesirable properties of coefficients.

Note that functionals operating on ∇a are not defined for the weak assumptions on $\mathcal{A}$ of Definition 1.4, but can be replaced by difference quotients after discretization of the equations.

Adding a regularization functional as defined above, commonly referred to as *Tikhonov regularization*, is not the only possible method of regularization, although it is used in the vast majority of publications on parameter identification. See the book by Engl, Hanke, and Neubauer [32] for an overview of methods.

Using the definitions above, Problem 1.1 can be stated as follows:

Problem 1.7 (Continuous problem). *Minimize the regularized deviation*

$$J(u,a) = m(u-z) + \beta r(a)$$

of u *from the measurement* z *, with* $\beta \geq 0$ *being a regularization parameter, subject to the constraints:*

$$\begin{aligned} (a\nabla u, \nabla\varphi) - (f, \varphi) &= 0 \qquad \forall \varphi \in V_0, \\ u|_{\Gamma_D} &= g, \\ a_0 \leq a &\leq a_1. \end{aligned}$$

Solvability and uniqueness for this problem crucially depend on the exact form of the functionals $m(\cdot)$ and $r(\cdot)$, and the function spaces on which they operate. These questions are touched briefly in Section 1.8.

Before going on with the discussion of methods for solving the constrained optimization problem 1.7, we would like to point out that the constraints are of very different nature:

- *The state equation:* Since we expect to find the unknown parameter only approximately, it would be useless to require u to satisfy the state equation exactly in every step of the process.

- *Dirichlet boundary conditions:* Being linear, these can be observed exactly by setting the initial iterate u_0 such that it satisfies the boundary conditions exactly, and then take all updates δu from the linear subspace that has zero boundary conditions on Γ_D.

- *Bounds:* The lower bound $0 < \alpha \leq a$ needs to be satisfied exactly, since it guarantees well-posedness and solvability of the problem and also contains essential physical meaning. The actual bounds $a_0 \leq a \leq a_1$ may be violated slightly but their enforcement stabilizes the process, see Chapter 3.

It must be stressed that Problem 1.7 is only one possible formulation of the problem of parameter estimation. It has, among many other examples, been used very successfully for parameter identification and optimization problems in ODE and DAE systems by Bock et al. [22, 23, 57, 29], Schulz [58], and Becker et al. [16, 18]. Haber and Oldenburg [38] use it for applications in parameter estimation problems involving elliptic partial differential equations. However, there are many other possible formulations. For example, it is common practice in applied sciences to treat the state variable as dependent on the parameter, thus eliminating the explicit state equation constraint, see for example Kravaris and Seinfeld [47] and Haber et al. [38]. The resulting formulation is often referred to as *Output Least Squares (OLS)* because it tries to minimize the square of the difference between measurement and the output of the differential equation operator for a given set of parameters. Furthermore, the state equation constraint can be treated using a primal-dual strategy (Bergounioux et al. [20]), or using an augmented Lagrangian approach (Kunisch et al. [42]). For further possible duality methods, see for example Chavent et al. [24, 26].

1.2 Optimality conditions and stability

In the following sections, we will develop an approach to solve the constrained minimization problem 1.7 by using a Lagrangian formulation and Newton's method. In a first step, we state the necessary conditions for an optimum in this section, and prove stability of solutions under suitable conditions. We then discuss second order conditions, and finally show the first order conditions for the constrained problem. For the time being, we defer the inclusion of the bound constraints $a_0 \leq a(\mathbf{x}) \leq a_1$ to Chapter 3 and assume that they are fulfilled even if not explicitly included in the problem.

1.2.1 First order conditions

Assuming that the inequality constraints $a_0 \leq a(\mathbf{x}) \leq a_1$ are non-existent, or inactive at the solution, we formulate Problem 1.7 by introducing a Lagrange multiplier for the state equation constraint and searching for a stationary point of the corresponding Lagrangian functional.

Problem 1.8 (Unconstrained first order conditions). *Let $\lambda \in V_0(\Omega)$ be a Lagrange multiplier and let*

$$L(u, a, \lambda) = m(u - z) + \beta r(a) + (\nabla\lambda, a\nabla u) - (\lambda, f) \tag{1.2}$$

denote the Lagrangian of the problem, then the solution

$$x = \{u, a, \lambda\} \in \mathcal{X}_g = V_g \times \mathcal{A} \times V_0$$

of problem 1.7, with inequality constraints $a_0 \leq a \leq a_1$ neglected, is determined by the first order necessary conditions

$$\nabla_x L(x; y) = 0 \qquad \forall y = \{\varphi, \chi, \psi\} \in \mathcal{X}_0 = V_0 \times \mathcal{A} \times V_0. \tag{1.3}$$

In explicit form, equation (1.3) reads: *Find $x = \{u, a, \lambda\} \in \mathcal{X}_g$ such that for all $y = \{\varphi, \chi, \psi\} \in \mathcal{X}_0$*

$$\nabla_u L(x; \varphi) \equiv m'(u - z; \varphi) + (\nabla\lambda, a\nabla\varphi) = 0, \tag{1.4}$$

$$\nabla_a L(x; \chi) \equiv \beta r'(a; \chi) + (\nabla\lambda, \chi\nabla u) = 0, \tag{1.5}$$

$$\nabla_\lambda L(x; \psi) \equiv (\nabla\psi, a\nabla u) - (\psi, f) = 0. \tag{1.6}$$

The validity of the characterization of solutions of (1.3) relies on the existence of a Lagrange multiplier λ. This is proven, for example, in Ito and Kunisch [42].

1.2.2 Stability of solutions

Existence and uniqueness of solutions can be based on stability. In the following, we first show inf-sup stability for the simpler case that we are looking for a single scalar parameter only, and afterwards show it for the general case for a subset of parameters satisfying some smoothness property. Due to this latter restriction,

the result cannot be used to prove existence and uniqueness, but nevertheless reveals the dependence of solutions on perturbations in the data.

The first proposition proves stability for the case that we are trying to identify a constant parameter. For the proof, we require the existence of a regularization term, which seems unnecessary for this simple case. We nevertheless state this case as it sets the stage for the following proof concerning distributed coefficients, but note that we consider it likely that the inf-sup constant can be made independent of the regularization parameter.

Proposition 1.9 (Stability for constant parameters). *Assume we want to identify a constant parameter $a \in \mathbb{R}$. Let $m(u-z) = \frac{1}{2}\|\nabla(u-z)\|^2$, $r(a) = \frac{1}{2}|a|^2$, and assume for simplicity that u has zero boundary values. Then the solution $x = \{u, a, \lambda\} \in \mathcal{X}_0 = H_0^1 \times \mathbb{R} \times H_0^1$ of (1.3) satisfies the system*

$$A(x, y) = (\nabla z, \nabla\varphi) + (f, \psi) \qquad \forall y = \{\varphi, \chi, \psi\} \in \mathcal{X}_0,$$

arising from (1.3) by reordering of terms, with the semilinear form defined as

$$A(x, y) = (\nabla u, \nabla\varphi) + (\nabla\lambda, \nabla\varphi)a + (\nabla u, \nabla\psi)a + \beta a\chi + (\nabla u, \nabla\lambda)\chi.$$

Then with $\|x\|_{\mathcal{X}}^2 = \|\nabla u\|_{L^2}^2 + |a|^2 + \|\nabla\lambda\|_{L^2}^2$ there exists $\gamma > 0$ such that the inf-sup condition

$$\sup_{y \in \mathcal{X}_0} \frac{A(x, y)}{\|y\|_{\mathcal{X}}} \geq \gamma \|x\|_{\mathcal{X}},$$

holds for all $x = \{u, a, \lambda\} \in \mathcal{X}_0$ satisfying $0 < a_0 \leq a < \infty$.

Proof. For each $x = \{u, a, \lambda\}$, we choose a test function $\hat{y} = \{\lambda, \frac{1}{\beta}a^2, u - (\frac{1}{a} + \frac{a}{\beta})\lambda\}$ such that first we have

$$A(x, \hat{y}) = a\|x\|_{\mathcal{X}}^2$$

by cancellation of the cross-terms $(\nabla u, \nabla\lambda)$. On the other hand, $\hat{y}$ is chosen in such a way that we can bound $\|\hat{y}\|_{\mathcal{X}}$ by $\|x\|_{\mathcal{X}}$, by absorbing the cross-term into the norms of u, λ, and choosing the factors such that the components of $\|\hat{y}\|_{\mathcal{X}}$ are balanced. To see this, we compute the norm of $\hat{y}$:

$$\|\hat{y}\|_{\mathcal{X}}^2 = \|\nabla u\|^2 + \left(\frac{a}{\beta}\right)^2 |a|^2 + \left[1 + \left(\tfrac{1}{a} + \tfrac{a}{\beta}\right)^2\right] \|\nabla\lambda\|^2 - 2\left(\tfrac{1}{a} + \tfrac{a}{\beta}\right)(\nabla u, \nabla\lambda).$$

Using Young's inequality and comparing the relative sizes of the factors in front of the norms of the components of x, we then have

$$\|\hat{y}\|_{\mathcal{X}}^2 \leq \left[\tfrac{3}{4} + \left(\tfrac{1}{2} + \tfrac{1}{a} + \tfrac{a}{\beta}\right)^2\right] \|x\|_{\mathcal{X}}^2.$$

Thus,

$$\sup_{y \in \mathcal{X}_0} \frac{A(x, y)}{\|y\|_{\mathcal{X}}} \geq \frac{A(x, \hat{y})}{\|\hat{y}\|_{\mathcal{X}}} \geq \gamma \|x\|_{\mathcal{X}}$$

with

$$\gamma = \min_{a_0 \le a < \infty} \frac{a\beta}{\sqrt{\frac{3}{4}\beta^2 + \left(\frac{\beta}{2} + \frac{\beta}{a} + a\right)^2}} = \frac{a_0\beta}{\sqrt{\frac{3}{4}\beta^2 + \left(\frac{\beta}{2} + \frac{\beta}{a_0} + a_0\right)^2}}.$$

□

The proof carries over directly to the case of discretized state and adjoint variable.

It is not possible to extend the proof of the theorem to the case of a distributed coefficient in a simple way, since then the choice of a test function $\hat{y}$ depending on x in a nonlinear way is not possible any more. However, the following result holds:

Theorem 1.10 (Stability for the distributed case). *For the case of a distributed coefficient, let $\tilde{\mathcal{A}} \subset \mathcal{A}$ be the set of functions $a \in \mathcal{A}$ satisfying the bound $a \ge a_0$ almost everywhere and for which we can find functions $\bar{\alpha}$ which satisfy the smoothness condition*

$$\sup_{\varphi \in H_0^1} \frac{\left\|\nabla\varphi - a\nabla(\frac{1}{\bar{\alpha}}\varphi)\right\|}{\|\nabla\varphi\|} \le \varepsilon < a_0, \tag{1.7}$$

and for some constant M the condition

$$\left\|\frac{1}{\bar{\alpha}} + \frac{a_0}{\beta}\right\|_{W^{1,\infty}} \le M < \infty. \tag{1.8}$$

Then there exists $\gamma > 0$ such that the inf-sup condition

$$\sup_{y \in \mathcal{X}_0} \frac{A(x,y)}{\|y\|_{\mathcal{X}}} \ge \gamma \|x\|_{\mathcal{X}},$$

holds for all $x \in H_0^1 \times \tilde{\mathcal{A}} \times H_0^1$, where

$$A(x,y) = (\nabla u, \nabla\varphi) + (a\nabla\lambda, \nabla\varphi) + (a\nabla u, \nabla\psi) + \beta(a,\chi) + (\nabla u \cdot \nabla\lambda, \chi),$$

and $\|x\|_{\mathcal{X}}^2 = \|\nabla u\|_{L^2}^2 + \|a\|_{L^2}^2 + \|\nabla\lambda\|_{L^2}^2$.

Proof. The proof follows the same ideas as that of Proposition 1.9. However, since the coefficient is no more a scalar, we can't use factors of it in the test functions, since we will have to take gradients of it. Rather, we use a smoothed version $\bar{\alpha}$ of the coefficient a as factor for u and λ.

For the proof, we consider for each given x the special test function $\hat{y} = \{\lambda, \frac{a_0-\varepsilon}{\beta}a, u - (\frac{1}{\bar{\alpha}} + \frac{a_0-\varepsilon}{\beta})\lambda\}$. Then,

$$A(x,\hat{y}) = (a\nabla u, \nabla u) + (a\nabla\lambda, \nabla\lambda) + (a_0 - \varepsilon)\|a\|^2 + \left(\nabla u, \nabla\lambda - a\nabla(\tfrac{1}{\bar{\alpha}}\lambda)\right).$$

Using the bound $a \ge a_0$ in the first two terms and condition (1.7) for the last term, we have

$$\begin{aligned} A(x,\hat{y}) &\ge a_0\|\nabla u\|^2 + a_0\|\nabla\lambda\|^2 + (a_0 - \varepsilon)\|a\|^2 - \varepsilon\|\nabla u\|\,\|\nabla\lambda\| \\ &\ge (a_0 - \varepsilon)\|x\|_{\mathcal{X}}^2. \end{aligned}$$

By assumption, the factor $a_0 - \epsilon$ is positive.

On the other hand, let $\omega = \frac{1}{\bar{\alpha}} + \frac{a_0-\varepsilon}{\beta}$. Then

$$\|\hat{y}\|^2 = \|\nabla u\|^2 + \|\nabla\lambda\|^2 + \left(\frac{a_0-\varepsilon}{\beta}\right)^2 \|a\|^2 + 2\,(\nabla u, \nabla(\omega\lambda)) + \|\nabla(\omega\lambda)\|^2.$$

We estimate $\|\nabla(\omega\lambda)\|$ by using the boundedness of ω in $W^{1,\infty}$ due to assumption (1.8), and by Poincaré's inequality on the norm of $\lambda \in H_0^1$, to obtain

$$\|\nabla(\omega\lambda)\| \le \|\omega\|_\infty \|\nabla\lambda\| + \|\nabla\omega\|_\infty \|\lambda\| \le C_\Omega \|\omega\|_{W^{1,\infty}} \|\nabla\lambda\| = C_\Omega M \|\nabla\lambda\|.$$

Thus,

$$\begin{aligned}
\|\hat{y}\|^2 &\le \|\nabla u\|^2 + \|\nabla\lambda\|^2 + \left(\frac{a_0-\varepsilon}{\beta}\right)^2 \|a\|^2 \\
&\qquad + 2C_\Omega M \|\nabla u\|\|\nabla\lambda\| + C_\Omega^2 M^2 \|\nabla\lambda\|^2, \\
&\le (1 + C_\Omega M)\|\nabla u\|^2 + (1 + C_\Omega M + C_\Omega^2 M^2)\|\nabla\lambda\|^2 + \left(\frac{a_0-\varepsilon}{\beta}\right)^2 \|a\|^2 \\
&\le \max\left\{1 + C_\Omega M + C_\Omega^2 M^2, \left(\frac{a_0-\varepsilon}{\beta}\right)^2\right\} \|x\|_{\mathcal{X}}^2,
\end{aligned}$$

and the claimed result holds with

$$\gamma = \frac{a_0 - \varepsilon}{\max\left\{\sqrt{\frac{3}{4} + (\frac{1}{2} + C_\Omega M)^2}, \left(\frac{a_0-\varepsilon}{\beta}\right)\right\}}.$$

□

Theorem 1.10 shows that the stability properties of solutions deteriorate as expected if the amount of regularization is reduced, since $\gamma < \beta$. On the other hand, for fixed β, the result shows that physically meaningful solutions satisfying the condition on the parameter are stable if a_0 is sufficiently large.

Remark 1.11. *The requirement* (1.7) *on the elements of* $\tilde{\mathcal{A}}$ *can be rewritten as follows: for each* $a \in \tilde{\mathcal{A}}$ *there must be a function* $\bar{\alpha}$ *satisfying*

$$\sup_{\varphi \in H_0^1} \frac{\left\|(1 - \frac{a}{\bar{\alpha}})\nabla\varphi + \varphi \frac{a}{\bar{\alpha}} \frac{\nabla\bar{\alpha}}{\bar{\alpha}}\right\|}{\|\nabla\varphi\|} \le \varepsilon < a_0.$$

Using Poincaré's inequality on $\varphi \in H_0^1$, *this condition is satisfied if we can find an approximation* $\bar{\alpha}$ *to* a *such that*

$$\left\|1 - \frac{a}{\bar{\alpha}}\right\| + C_\Omega \left\|\frac{a}{\bar{\alpha}} \frac{\nabla\bar{\alpha}}{\bar{\alpha}}\right\| \le \varepsilon < a_0.$$

This implies closeness of $\bar{\alpha}$ *to* a *as well as smallness of* $\nabla\bar{\alpha}$. *The theorem shows that the stability deteriorates as* ε *grows.*

If we are looking for Lipschitz continuous coefficients, then the condition is satisfied if $a \geq a_0 > 0$ and $\|\nabla a\| \leq \varepsilon a_0 < a_0^2$, by choosing $\bar{\alpha} = a$. For constant coefficients, we have that $\varepsilon = 0, A = \frac{a_0}{\beta}$, and we can recover the result of Proposition 1.9, but with γ worse by a constant factor of C_Ω.

Remark 1.12. *Theorem 1.10 still holds if we replace the L^2-norm on $\mathcal{A}$ by any other norm, if the regularization term is chosen accordingly. For example, the theorem holds if $\|x\|_{\mathcal{X}}^2 = \|\nabla u\|_{L^2}^2 + \|a\|_{H^1}^2 + \|\nabla\lambda\|_{L^2}^2$ and $r(a) = \frac{1}{2}\|a\|_{H^1}^2$.*

1.2.3 Second order conditions

As for finite dimensional problems, the second order necessary conditions for an optimum $\{u, a\}$ are that

$$\nabla^2_{\{u,a\}} L(x; \{\delta u, \delta a\}, \{\delta u, \delta a\}) > 0 \tag{1.9}$$

holds for all directions $\{\delta u, \delta a\}$ tangential at x to the feasible set defined by $-\nabla\cdot(a\nabla u) = f$, i.e. for all $\delta u, \delta a$ satisfying

$$-\nabla\cdot(a\nabla\delta u) - \nabla\cdot(\delta a\nabla u) = 0,$$

see, e.g., Maurer and Zowe [50].

For a special, although slightly unrealistic, choice of measurement and regularization functionals, it is simple to show that these conditions always hold for an optimum of Problem 1.8 if measurement noise is small enough, or is countered by a sufficiently large regularization parameter:

Proposition 1.13. *Assume $m(\varphi) = \frac{1}{2}\|\nabla\varphi\|^2$, $r(\chi) = \frac{1}{2}\|\chi\|_k^2, k > \dim\Omega/2$. Assume further that Ω is a bounded domain with Lipschitz continuous boundary, and that at the solution $x = \{u, a, \lambda\}$ the misfit is $m(u - z) < \varepsilon$. Then the second order necessary optimality conditions for the Hessian* (1.9) *hold for all perturbations $\delta u \in H_0^1, \delta a \in H^k$.*

Proof. By assumed continuity, convexity, and positivity of $m(\cdot)$, we infer from $m(u - z) < \varepsilon$ that there exists $\delta > 0$, $\lim_{\varepsilon\to 0}\delta(\varepsilon) = 0$, growing strictly monotonously with ϵ such that $\|m'(u - z; \cdot)\|_{H^{-1}} < \delta$. Due to (1.4) and using standard elliptic estimates, we therefore have $\|\lambda\|_{H^1} < \delta/a_0$.

On the other hand, by convexity of $m(\cdot)$ and $r(\cdot)$, there are constants $\mu > 0, \rho > 0$ with $\mu = \inf_{\delta u} m''(\delta u, \delta u)/\|\delta u\|_{H^1}^2$, $\rho = \inf_{\delta a} r''(a; \delta a, \delta a)/\|\delta a\|_k^2$. Finally, using the definition of the Lagrangian, the condition reads

$$\begin{aligned}
\nabla^2_{\{u,a\}} L(x; \{\delta u, \delta a\}, \{\delta u, \delta a\}) &= m''(\delta u, \delta u) + \beta r''(a; \delta a, \delta a) + (\nabla\lambda, \delta a\nabla\delta u) \\
&\geq \mu\|\delta u\|_{H^1}^2 + \beta\rho\|\delta a\|_k^2 - \|\lambda\|_{H^1}\|\delta a\|_{0,\infty}\|\delta u\|_{H^1} \\
&\geq \mu\|\delta u\|_{H^1}^2 + \beta\rho\|\delta a\|_k^2 - \tfrac{C_e\delta}{a_0}\|\delta a\|_k\|\delta u\|_{H^1},
\end{aligned}$$

where in the last step we have made use of the Sobolov inequality $\|\delta a\|_{0,\infty} \leq C_e\|\delta a\|_k$ that holds for the chosen class of domains Ω. Thus, if β large enough, or δ and thus ε small enough, the entire term is larger than zero and the second order condition holds. □

The result shows that large noise may lead to irregular points in the Lagrangian unless it is countered by an increased regularization parameter $\beta > C_e^2\delta^2/(4a_0^2\mu\rho)$. Note that this then implies stability of the solution a with respect to perturbations in the measurements z. However, as in the stability theorems above, the stability constant is only proportional to β. For practical purposes, the proposition above is rather uninteresting, since the regularization functional has to be chosen too strong.

1.2.4 First order conditions for the constrained problem

Previously, we have assumed that inequality constraints $a_0 \le a \le a_1$ either do not exist or are inactive. Although we will base the rest of this chapter on this assumption and present their inclusion into the numerical procedure only in Chapter 3, we state the first order conditions of the bound constrained problem for completeness. For this, let us first define the cone $\mathcal{C}$ and dual cone $\mathcal{C}^+$

$$\mathcal{C} = \{a \in L^\infty : a \ge 0\}, \qquad \mathcal{C}^+ = \{\chi \in L^1 : \langle \chi, a\rangle \le 0 \ \forall a \in \mathcal{C}\}. \tag{1.10}$$

Then, the constrained continuous problem can be stated in the following form:

Problem 1.14 (Constrained first order conditions). *Let $\lambda \in V_0$ and $\mu_i \in \mathcal{C}^+, i = 1, 2$, be Lagrange multipliers for the state equation and lower and upper bounds, respectively, and let*

$$\begin{aligned} L(u, a, \lambda, \mu_0, \mu_1) = m(u - z) + \beta r(a) + (\nabla\lambda, a\nabla u) - (\lambda, f) \\ + (\mu_0, a - a_0) + (\mu_1, a_1 - a) \end{aligned} \tag{1.11}$$

denote the Lagrangian of the problem, then the solution $x = \{u, a, \lambda, \mu_0, \mu_1\}$ of Problem 1.7 is determined by the first order necessary condition

$$\begin{aligned} \nabla_{\{u,a,\lambda\}} L(x; y) &= 0 && \forall y = \{\varphi, \chi, \psi\} \in \mathcal{X}_0 = V_0 \times \mathcal{A}'[a] \times V_0, \\ \nabla_{\mu_i} L(x; \gamma) &\le 0 && \forall \gamma \in \mathcal{C}^+, i = 1, 2, \\ (\mu_0, a - a_0)_{L^2} &= 0, \\ (\mu_1, a_1 - a)_{L^2} &= 0. \end{aligned}$$

A proof of this under slightly different conditions can be found in Ito and Kunisch [42].

1.3 Newton's method for the optimality conditions

Due to their nonlinearity, a direct solution of the first order conditions (1.3) is not possible; we therefore employ a Newton iteration to generate a sequence of iterates $x_k = \{u_k, a_k, \lambda_k\}$ hopefully converging to the exact solution $x = \{u, a, \lambda\}$ of (1.3) as $k \to \infty$. The treatment of bound constraints $a_0 \le a \le a_1$ will later be included into the computation of Newton steps, but we defer this to Chapter 3.

Newton's method, as applied here, consists of two steps: first compute a search direction δx_k in which the updates for x_k to get to x_{k+1} will be chosen. Then, the length of the step in this direction is chosen. These two steps will be discussed in the following. We note that the approach chosen here is fully equivalent to the Sequential Quadratic Programming (SQP) method as long as bound constraints are not incorporated.

Conceptually, the method proposed here can be described either on a continuous or a discrete level: either we fix a discretization and apply a number of Newton steps until we are satisfied with the convergence on this mesh; we then repeat the same steps on a finer discretization, of course using the old solution as a starting value. Or, alternatively, we consider the steps on a continuous level and compute an approximation of the continuous search direction by separately discretizing each step, using a priori unrelated discretizations; in practice, discretizations will be changed after a few steps if we are satisfied with the reduction of the residual on this mesh.

Although formally equivalent, we prefer to view the algorithm the second way. We then have an iteration in infinite dimensional function spaces, which is more natural since we are interested in the solution of the problem in these spaces, rather than on any arbitrarily chosen fixed discretization. The residual of the optimality condition is thus measured in continuous norms, and errors are computed with respect to the continuous solution. Also, the discussion of a stopping criterion for iteration on a fixed mesh is replaced by a criterion for choosing a different discretization for the next Newton step.

Accordingly, the following discussion of Newton's method will be based on a purely continuous level, with discretization of each step being treated in the next section.

Computing the Newton search direction. In each step, Newton's method computes the next search direction by using a local approximation of the function which we want to find a zero of, i.e., of $\nabla_x L$. This is done by fitting a quadratic approximation to L, and taking the direction to the saddle point of this quadratic approximation as next search direction.

The conditions determining this search direction $\delta x_k = \{\delta u_k, \delta a_k, \delta\lambda_k\} \in \mathcal{X}_0$ are then the following equations:

$$\nabla_x^2 L(x_k; \delta x_k, y) = -\nabla_x L(x_k; y) \tag{1.12}$$

for all test functions $y = \{\varphi, \chi, \psi\} \in \mathcal{X}_0$, or explicitly:

$$\begin{aligned}
&m''(u_k - z; \delta u_k, \varphi) + (\nabla\lambda_k, \delta a_k\, \nabla\varphi) + (\nabla\delta\lambda_k, a_k\nabla\varphi) &&= -\nabla_u L(x_k;\varphi),\\
&(\nabla\lambda_k, \chi\nabla\delta u_k) \qquad + \beta r''(a_k; \delta a_k, \chi) + (\nabla\delta\lambda_k, \chi\nabla u_k) &&= -\nabla_a L(x_k;\chi),\\
&(\nabla\psi, a_k\nabla\delta u_k) \qquad + (\nabla\psi, \delta a_k\nabla u_k) &&= -\nabla_\lambda L(x_k;\psi).
\end{aligned} \tag{1.13}$$

The Gauß-Newton method. From the first order conditions (1.4) we see that λ is small whenever $m'(u - z; \cdot)$ is small. This holds at least near the solution, if the model (i.e. the state equation) chosen to describe the measurements z is correct, and if z does not contain too much noise. For the problems treated

in this work, we assume that these conditions are satisfied; such problems are termed *small residual problems.*

It is then a common simplification to omit the terms containing λ from the Hessian in the Newton step, resulting in the following equations instead of (1.13):

$$\begin{aligned} m''(u_k - z; \delta u_k, \varphi) \qquad\qquad + (\nabla \delta\lambda_k, a_k \nabla\varphi) &= -\nabla_u L(x_k; \varphi), \\ \beta r''(a_k; \delta a_k, \chi) + (\nabla \delta\lambda_k, \chi \nabla u_k) &= -\nabla_a L(x_k; \chi) \\ (\nabla\psi, a_k \nabla \delta u_k) \quad + (\nabla\psi, \delta a_k \nabla u_k) &= -\nabla_\lambda L(x_k; \psi). \end{aligned} \tag{1.14}$$

The resulting methods are called *Gauß-Newton methods* and have found very successful applications in parameter estimation and optimization (see, e.g., Bock et al. [22, 23], Schulz [58], or Pratt et al. [55]). This modification makes the problem to be solved in each iteration simpler, since the Schur complement with respect to the regularization block becomes positive definite under suitable conditions (see Lemma 1.21), while the original problem will be indefinite usually. Also, the computation of the Schur complement is simpler.

For the problems considered in this work, the pure Newton and Gauß-Newton methods perform equally well when comparing the number of iterations necessary for a certain accuracy. We have usually used the latter, in view of the simplifications occurring and in particular considering the size of the problems to be treated in Chapters 4 and 5. A comprehensive comparison of the suitability of Newton and Gauß-Newton search directions in parameter estimation problems can be found in Bock [23].

Computing the step length. Once the search direction is known, the second part of a safeguarded Newton method is to determine the step length α_k, by which we define the next iterate as $x_{k+1} = x_k + \alpha_k \delta x_k$. This is necessary since in practice the quadratic approximation of the Lagrangian is not an accurate description of the true behavior, except in the vicinity of x_k. Thus, safeguarding the length of a step in direction δx_k is necessary.

To compute a step length α_k, several methods are in common use, for example using the Goldstein-Armijo conditions. In general, they choose α_k as an approximation of the minimizer α_k^* of some objective function $p(\alpha_k) = p(x_k + \alpha_k \delta x_k)$. For constrained problems, this penalty function has to include the minimization functional $J(\cdot)$ as well as an appropriately weighted norm of the residual of the constraint.

Since the construction of a suitable weight for the norm of the constraints is difficult if the constraints are partial differential equations, we choose to minimize the norm of the residual of the optimality condition $\nabla_x L = 0$ instead. The proper norm for this residual would by the norm of the dual space $\mathcal{X}'$ of $\mathcal{X}_g = V_g \times \mathcal{A} \times V_0$. Since this involves H^{-1} norms, it is impractical to evaluate. Therefore, we evaluate its discrete analogon, i.e. the norm on the dual $\mathcal{X}_h'$ of the discretization space $\mathcal{X}_h = V_h \times \mathcal{A}_h \times V_h$ to be defined in the next section. For this, the following representation holds:

Lemma 1.15. *Denote by $g = (g_u, g_a, g_\lambda)^T$ the discrete gradient of the Lagrangian $L(x)$, i.e.*

$$(g_u)_i = \nabla_u L(x;\varphi_i), \qquad (g_a)_i = \nabla_a L(x;\chi_i), \qquad (g_\lambda)_i = \nabla_\lambda L(x;\psi_i),$$

where $\varphi_i, \chi_i, \psi_i$ are sets of functions forming a basis of the discretization space $\mathcal{X}_{0,h} = V_{0,h} \times \mathcal{A}_h \times V_{0,h}$ to be defined in the next section. Then the following identity holds:

$$\|\nabla_x L(x;\cdot)\|^2_{\mathcal{X}_h'} \equiv \sup_{y_h \in \mathcal{X}_h} \frac{L(x;y_h)^2}{\|y_h\|^2_{\mathcal{X}}} = g_u^T A^{-1} g_u + g_a^T M^{-1} g_a + g_\lambda^T A^{-1} g_\lambda,$$

where A, M are Laplace and mass matrices, defined by $A_{ij} = (\nabla\varphi_i, \nabla\varphi_j)$, $M_{ij} = (\chi_i, \chi_j)$, respectively. Furthermore, there holds

$$\|\nabla_x L(x;\cdot)\|_{\mathcal{X}_h'} \le \|\nabla_x L(x;\cdot)\|_{\mathcal{X}'}.$$

Proof. The first part follows immediately from the definition of norms on dual spaces, using that $\mathcal{X}_h$ is finite dimensional. The second part is obvious since $\mathcal{X}_h \subset \mathcal{X}$. □

Since the evaluation of the $\mathcal{X}_h'$ norm only involves the inversion of two Laplace matrices and one mass matrix, it is roughly as expensive as one evaluation of the Schur complement of the Hessian, see Section 1.5 below, and is thus comparably cheap.

The following lemma states that this norm is a valid penalty functional:

Lemma 1.16. *Let*

$$p(\alpha) = \|\nabla_x L(x_k + \alpha\delta x_k;\cdot)\|^2_{\mathcal{X}_h'}.$$

Then full Newton search directions δx_k are directions of descent of p, i.e. $p'(0) < 0$.

Proof. As shown in Lemma 1.15, the norm on $\mathcal{X}_h'$ is induced by a scalar product. With $g(x_k + \alpha\delta x_k)$ the projection of $\nabla_x L(x_k + \alpha\delta x_k)$ as defined in Lemma 1.15, we have

$$\begin{aligned} p(\alpha) &= \|g(x_k + \alpha\delta x_k)\|^2_{[A^{-1},M^{-1},A^{-1}]} \\ &= \|g_u(x_k + \alpha\delta x_k)\|^2_{A^{-1}} + \|g_a(x_k + \alpha\delta x_k)\|^2_{M^{-1}} + \|g_\lambda(x_k + \alpha\delta x_k)\|^2_{A^{-1}}, \end{aligned}$$

with $\|v\|^2_B = v^T B v$. Then

$$p'(0) = 2\left\langle g(x_k), \tfrac{d}{d\alpha} g(x_k + \alpha\delta x_k)\big|_{\alpha=0} \right\rangle_{[A^{-1},M^{-1},A^{-1}]}$$

By definition of g and of the full Newton search direction δx_k, there holds

$$\tfrac{d}{d\alpha} g(x_k + \alpha\delta x_k)\big|_{\alpha=0} = -g(x_k),$$

and the claim follows by positive definiteness of A^{-1} and M^{-1}. □

If the quadratic approximation of the Lagrangian used for the Newton step were exact, then $p(\alpha)$ would be a quadratic function, and since $p'(0) = -2p(0)$, it would have its minimum at $\alpha = 1$, i.e. the resulting step length would be optimal. Numerical experiments indicate that comparably good step lengths can be obtained by replacing A^{-1} in the evaluation of the inverse norm by M^{-1}, which is significantly cheaper to evaluate. Even diagonal approximations of the matrices result in good step lengths, keeping in mind that step length selection is only an aid in finding the solution and that we are in general not interested in *optimal* step lengths.

1.4 Discretization of Newton steps

For actual computations, we need to discretize the problem. As discussed above, we do this separately for each Newton step. The choice of meshes and discrete spaces used here differs from common practice in the majority of the available literature in that the coefficient is discretized separately. In this section, we give a short definition of the finite element spaces we use, and then explain their use in the discretization and the connections to the meshes we use.

We start by briefly defining the usual piecewise polynomial spaces used in finite element methods:

Definition 1.17 (Spaces on unit cells). *Let $\hat{K}$ be the unit element $[0,1]^d$, i.e. the unit square in two and the unit cube in three space dimensions. Then the Lagrange interpolation space of order r on $\hat{K}$ is defined by*

$$\hat{Q}^r(\hat{K}) = \left\{ \varphi : \hat{K} \to \mathbb{R} \quad | \quad \varphi = \prod_{i=1}^{d} \sum_{j=0}^{r} c_{ij} x_i^j \right\}.$$

Definition 1.18 (Spaces on real cells). *Let K be an element of a mesh, such that there exists a (bi-, tri-)linear mapping $\Phi : \hat{K} \to K$ from the unit cell to the cell in real space. Then the Lagrange interpolation spaces are defined as follows:*

$$Q^r(K) = \left\{ \varphi(\mathbf{x}) : K \to \mathbb{R} \quad | \quad \exists \hat{\varphi}(\hat{\mathbf{x}}) \in \hat{Q}^r(\hat{K}), \varphi(\mathbf{x}) = \hat{\varphi}(\Phi^{-1}(\mathbf{x})) \right\}.$$

Definition 1.19 (Meshes). *Let the domains on which we consider partial differential equations in this thesis, be bounded open subsets Ω of $\mathbb{R}^d, d = 1, 2, 3$. Assume Ω is polygonal. A subdivision $\mathbb{T} = \{K\}$ is called a mesh in the context of this thesis if it satisfies the following properties:*

- *$K_i \cap K_j = \emptyset$, for $K_i, K_j \in \mathbb{T}, i \neq j$; $\quad \bigcup_K \overline{K} = \overline{\Omega}$;*
- *each cell $K \in \mathbb{T}$ is the image of the unit cell $\hat{K} = [0,1]^d$ under a polynomial mapping, i.e. the cells are lines, quadrilaterals, or hexahedra, depending on the space dimension.*

For various estimates, we also require the regularity condition that the eigenvalues of the Jacobian matrix of the mapping between unit cell $\hat{K}$ and real cells K are bounded from below and above.

Definition 1.20 (Spaces on meshes). *Let $\mathbb{T} = \{K\}$ be a mesh as defined above. Then the spaces of continuous functions of piecewise polynomials of degree r on $\mathbb{T}$ are defined by*

$$Q_c^r(\mathbb{T}) = \big\{\varphi : \Omega \to \mathbb{R} \quad | \quad \varphi \text{ continuous on } \Omega, \\ \varphi|_K \in Q^r(K) \quad \forall K \in \mathbb{T}\big\},$$

and the respective spaces of discontinuous functions are

$$Q_d^r(\mathbb{T}) = \big\{\varphi : \Omega \to \mathbb{R} \quad | \quad \varphi|_K \in Q^r(K) \quad \forall K \in \mathbb{T}\big\}.$$

With these definitions, we can discuss the function spaces and mesh types used in the discretization of the Newton steps:

Finite Element Spaces. Of central importance is the choice of the discrete finite element spaces $\mathcal{U}_h, \mathcal{A}_h, *_h$ for the primal variable u, the parameter a , and the adjoint variable λ. By symmetry of the formulation of the problem, it is reasonable to choose $*_h = \mathcal{U}_h$, and for $\mathcal{U}_h$ to take the usual piecewise tensor product polynomial function spaces $Q_c^r(\mathbb{T})$ of degree r on a given mesh $\mathbb{T}$.

Formally, we choose the following finite element spaces:

- for the discretized state and adjoint variables u_h, λ_h: $\mathcal{U}_h = *_h = Q_c^r(\mathbb{T})$, i.e., the spaces of globally continuous functions made up of piecewise tensor product polynomials of degree r over a mesh $\mathbb{T}$;
- for the discretized parameter a_h: $\mathcal{A}_h = Q_c^{r'}(\mathbb{T}_a)$ or $\mathcal{A}_h = Q_d^{r'}(\mathbb{T}_a)$, i.e., the spaces of continuous or discontinuous functions of piecewise polynomial degree r' over a mesh $\mathbb{T}_a$.

Choosing different spaces for $\mathcal{U}_h$ and $\mathcal{A}_h$ is an aspect in which this work deviates from usual practice in the literature. There, most often spaces of piecewise bilinear functions are used for both primal and dual variables, as well as the coefficient, mostly for convenience (almost all publications cited within this work fall into this category). Note however Banks and Kunisch [13, 3] for examples where different spaces are used although restricted to the use of fixed uniformly refined meshes in only one space dimension. See also Chavent and Bissell [25] and Ben Ameur et al. [19] for some experiments on choosing a discretization of the coefficient.

Meshes. In most of the available literature, not only the same finite element spaces for state, adjoint and parameter variable are used, seemingly all also use the same mesh. We pursue a more general approach by taking different, though related meshes for state and adjoint variable on the one hand and the parameter variable on the other hand.

This has advantages both on the analytical as well as on the numerical side:

- Choosing different meshes for state/adjoint and coefficient variables allows to resolve the local features of both independently.

- Choosing coarser meshes and lower-order function spaces for the coefficient acts as an additional regularization, since it reduces the possibilities for variation in the parameter. This is sometimes referred to by *regularization by discretization* (see Banks and Kunisch [13] and Kaltenbacher [46]), although this is usually meant in the context of fixed meshes. Choosing adaptive meshes allows for locally different amounts of regularization.

- Stability properties of the discretized saddle point problems are affected by the choice of discrete function spaces. Numerical experience indicates that it is beneficial to use a coarser mesh and/or lower order polynomials for the parameter variable.

- Choosing a coarser discretization for the coefficient can be understood as *adaptive model reduction*. This greatly reduces the numerical effort needed to compute solutions.

- We are anticipating extension to time dependent problems, where different meshes have to be chosen anyway: the mesh for the state variable changes with time, while the coefficient is usually constant in time. Furthermore, regularity levels of state variable and coefficient differ.

In this work, we will therefore use two meshes, $\mathbb{T}$ and $\mathbb{T}_a$, for state and adjoint variable, and the parameter, respectively. For implementational reasons, we require that $\mathbb{T}$ can be obtained from $\mathbb{T}_a$ by refinement. Taking $\mathbb{T}_a = \mathbb{T}$ is included as a special case.

1.5 The discretized problem

In each Newton step, the search direction is computed approximately by discretizing (1.12) using the spaces defined in the last section. Choosing bases $\{\varphi_i\}, \{\chi_i\}$ and $\{\psi_i\}$ and expanding the updates $\delta u, \delta a$ and $\delta\lambda$ with respect to these bases yields the following Karush-Kuhn-Tucker (KKT) matrix system:

$$\begin{pmatrix} M & B^T & A^T \\ B & R & C^T \\ A & C & 0 \end{pmatrix} \begin{pmatrix} \delta u_k \\ \delta a_k \\ \delta\lambda_k \end{pmatrix} = \begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix}. \tag{1.15}$$

The individual blocks in matrix and right hand side are defined by

$$M = \Big[\, m''(u_k - z; \varphi_i, \varphi_j) \,\Big]_{ij}, \quad F_1 = \Big[\, -m'(u_k - z; \varphi_i) - (a_k \nabla\lambda_k, \nabla\varphi_i) \,\Big]_i,$$
$$B = \Big[\, (\chi_i, \nabla\lambda \cdot \nabla\varphi_j) \,\Big]_{ij}, \qquad F_2 = \Big[\, -\beta r'(a_k; \chi_i) - (\nabla\lambda_k \cdot \nabla u_k, \chi_i) \,\Big]_i,$$
$$A = \Big[\, (a_k \nabla\varphi_i, \nabla\psi_j) \,\Big]_{ij}, \qquad F_3 = \Big[\, -(a_k \nabla u_k, \nabla\psi_i) + (f, \psi_i) \,\Big]_i,$$
$$R = \Big[\, \beta r''(a_k; \chi_i, \chi_j) \,\Big]_{ij}, \qquad C = \Big[\, (\nabla u_k \cdot \nabla\psi_i, \chi_j) \,\Big]_{ij},$$

where M corresponds to the misfit functional, R to the regularization functional, B and C to hyperbolic transport operators $\nabla\lambda \cdot \nabla + \Delta\lambda$ and $\nabla u \cdot \nabla$, and A is the matrix associated with the state equation.

By block elimination, (1.15) can be reformulated to yield a system where we first solve for δa_k, and only afterwards for δu_k and $\delta\lambda_k$. The equation for δa_k resulting from the full Newton equation has the form

$$\left\{R - \begin{bmatrix} B & C^T \end{bmatrix} \begin{bmatrix} 0 & A^{-1} \\ A^{-T} & -A^{-T}MA^{-1} \end{bmatrix} \begin{bmatrix} B^T \\ C \end{bmatrix}\right\} \delta a_k = F_2 - \begin{bmatrix} B & C^T \end{bmatrix} \begin{bmatrix} 0 & A^{-1} \\ A^{-T} & -A^{-T}MA^{-1} \end{bmatrix} \begin{bmatrix} F_1 \\ F_3 \end{bmatrix}, \quad (1.16)$$

where the system matrix on the left hand side is called the *Schur complement* of the KKT matrix (1.15) with respect to the R block. The updates for δu_k and $\delta\lambda_k$ are then obtained from

$$\begin{aligned} A\ \delta u_k &= F_3 - C\delta a_k, \\ A^T\delta\lambda_k &= F_1 - B^T\delta a_k - M\delta u_k. \end{aligned} \quad (1.17)$$

If we use the Gauß-Newton method, the block B in (1.15) is dropped, and the Schur complement solution requires the subsequent solution of the following three equations:

$$\begin{aligned} \left\{R + C^TA^{-T}MA^{-1}C\right\}\delta a_k &= F_2 - C^TA^{-T}F_1 + C^TA^{-T}MA^{-1}F_3, \\ A\ \delta u_k &= F_3 - C\delta a_k, \\ A^T\delta\lambda_k &= F_1 - M\delta u_k. \end{aligned} \quad (1.18)$$

1.6 Condition numbers of the linear problems

The choice of solvers for the linear problems to be solved in each Newton step crucially depends on the condition number of the Newton and Schur complement matrices. Fig. 1.1 shows a typical eigenvalue distribution of these matrices. Table 1.1 displays the eigenvalues of minimal and maximal absolute value of a sequence of Newton matrices, along with the condition number in the spectral norm. The condition number of the whole Newton matrix grows as h^{-6}, for the L^2 misfit minimization, and h^{-4} for H^1 minimization. The condition number of the whole matrix is not significantly changed by dropping the B block in the Gauß-Newton method and does also not vary much as iterations proceed on one mesh.

Contrary to this, Table 1.2 shows that the condition number of the Schur complement matrices is $\mathcal{O}(h^{-4})$ and $\mathcal{O}(h^{-2})$, depending on the choice of the misfit functional, and thus by two orders better than that of the full Newton matrix (this has previously been observed in Ascher and Haber [4]).

1.7 Solution of the linear problems

For the solution of the linear systems (1.15) arising in each Newton step, several methods have been tested. The most successful, robust, and extensible

Figure 1.1: *Left: Spectrum of the whole Newton matrix (left) and its Schur complement (right) for a typical discretization with* 81 *degrees of freedom for* u_h *and* λ_h *each, and* 16 *degrees of freedom for* a_h*. The condition numbers are* $\kappa \approx 1.5{\cdot}10^5$ *and* $\kappa \approx 600$*, respectively.*

	$m(u-z) = \frac{1}{2}\|u-z\|_\Omega^2$			$m(u-z) = \frac{1}{2}\|\nabla(u-z)\|_\Omega^2$		
h	$\min\|\mu_i\|$	$\max\|\mu_i\|$	κ_2	$\min\|\mu_i\|$	$\max\|\mu_i\|$	κ_2
2^{-3}	$5.06{\cdot}10^{-5}$	7.62	$1.5{\cdot}10^5$	$6.24{\cdot}10^{-3}$	9.74	$1.6{\cdot}10^3$
2^{-4}	$7.84{\cdot}10^{-7}$	7.90	$1.0{\cdot}10^7$	$3.96{\cdot}10^{-4}$	10.1	$2.6{\cdot}10^4$
2^{-5}	$1.22{\cdot}10^{-8}$	7.98	$6.5{\cdot}10^8$	$2.49{\cdot}10^{-5}$	10.2	$4.1{\cdot}10^5$
2^{-6}	$1.91{\cdot}10^{-10}$	7.99	$4.2{\cdot}10^{10}$	$1.56{\cdot}10^{-6}$	10.2	$6.6{\cdot}10^6$
2^{-7}	$2.99{\cdot}10^{-11}$	8.00	$2.7{\cdot}10^{12}$	$9.74{\cdot}10^{-8}$	10.2	$1.1{\cdot}10^8$
	$\mathcal{O}(h^6)$	$\mathcal{O}(1)$	$\mathcal{O}(h^{-6})$	$\mathcal{O}(h^4)$	$\mathcal{O}(1)$	$\mathcal{O}(h^{-4})$

Table 1.1: *Minimal and maximal eigenvalues* μ_i*, and condition number with respect to the spectral norm for the whole Newton matrix for two different misfit functionals* $m(\cdot)$*. The discretization is as in Fig. 1.1 (which corresponds to* $h = 2^{-3}$*). The mesh for* $h = 2^{-7}$ *has roughly 50,000 degrees of freedom.*

	$m(u-z) = \frac{1}{2}\|u-z\|_\Omega^2$			$m(u-z) = \frac{1}{2}\|\nabla(u-z)\|_\Omega^2$		
h	$\min\|\mu_i\|$	$\max\|\mu_i\|$	κ_2	$\min\|\mu_i\|$	$\max\|\mu_i\|$	κ_2
2^{-3}	$5.06{\cdot}10^{-5}$	$3.03{\cdot}10^{-2}$	$6.0{\cdot}10^2$	$6.26{\cdot}10^{-3}$	$1.70{\cdot}10^{-1}$	27
2^{-4}	$7.84{\cdot}10^{-7}$	$8.14{\cdot}10^{-3}$	$1.0{\cdot}10^4$	$3.96{\cdot}10^{-4}$	$5.27{\cdot}10^{-2}$	130
2^{-5}	$1.22{\cdot}10^{-8}$	$2.08{\cdot}10^{-3}$	$1.7{\cdot}10^5$	$2.49{\cdot}10^{-5}$	$1.48{\cdot}10^{-2}$	590
2^{-6}	$1.91{\cdot}10^{-10}$	$5.21{\cdot}10^{-4}$	$2.7{\cdot}10^6$	$1.56{\cdot}10^{-6}$	$4.35{\cdot}10^{-3}$	2800
2^{-7}	$2.99{\cdot}10^{-12}$	$1.31{\cdot}10^{-4}$	$4.4{\cdot}10^7$	$9.74{\cdot}10^{-8}$	$1.31{\cdot}10^{-3}$	13000
	$\mathcal{O}(h^6)$	$\mathcal{O}(h^2)$	$\mathcal{O}(h^{-4})$	$\mathcal{O}(h^4)$	$\mathcal{O}(h^2)$	$\mathcal{O}(h^{-2})$

Table 1.2: *Minimal and maximal eigenvalues* μ_i*, and condition number with respect to the spectral norm for the Schur complements of the same matrices as in Table 1.1. Note that the minimal eigenvalues are identical to those of the full Newton matrix.*

approach was solving the Schur complement form (1.18), when using the Gauß-Newton modification. We will describe this approach first. Other, less successful methods have also been tried, and will be discussed briefly afterwards.

1.7.1 Schur complement methods

Schur complement methods are known to be very efficient in many cases (see Schulz [58] for an overview of some Schur complement methods for optimization problems, or Turek [65] for flow problems). Since the Schur complement of the full Newton matrix (1.16) is too complicated for practical purposes, we invert the Gauß-Newton Schur complement (1.18) instead. This system may be solved by a Krylov space method for the (small) Schur complement, and a standard method to invert the Laplace matrices in each iteration.

The Schur complement matrix is not known explicitly, as A^{-1} and A^{-T} are only defined implicitly by solving a linear system with a specified right hand side. Thus, unless one wants to recover it by forming n matrix vector multiplications with it, we can only use iterative methods to invert the Schur complement matrix.

However, unlike the full Gauß-Newton matrix, the following lemma shows that the Schur complement is symmetric positive definite under reasonable conditions. We can then use the Conjugate Gradient method with its good convergence properties. By standard arguments, we have the following lemma:

Lemma 1.21 (Properties of Gauß-Newton Schur complement). *If the matrix R is symmetric positive definite and M symmetric and at least positive semidefinite, or if R is symmetric positive semidefinite and M is symmetric positive definite and C has full column rank, then the Schur complement matrix $R + C^T A^{-T} M A^{-1} C$ is symmetric positive definite.*

It is obvious that for the second case, the condition that M has to be positive definite can be replaced by the condition that it must be positive definite on the subspace $Y = \{y : y = A^{-1}Cx, x \in N(R)\}$, where $N(R)$ denotes the null space of R. However, it is difficult to characterize Y in order to check whether M is positive on it, in particular since it implicitly depends on the present iterates u_k, a_k through A and C.

The requirements stated in the lemma are what can usually be expected: the symmetry of M and R is given by the symmetry of second derivatives; their positive semidefiniteness is given by the assumed convexity of the functionals $m(\cdot)$ and $r(\cdot)$. Positive definiteness can be achieved, for example, by choosing one of the two to be a norm. The condition on C in the second possibility of the lemma can be shown to be equivalent to the condition that u_k must not be constant on certain patches of cells; as this can hardly be guaranteed in practice, it is better to choose R positive definite.

Note that when using the full Newton system, i.e. without the Gauß-Newton modification, then the Schur complement is symmetric but may not be positive definite. We are then forced to use a more expensive method than CG. Also, multiplications with the Schur complement of the full Newton matrix take four

instead of two multiplications with A^{-1} or A^{-T}, making the iterative solution significantly more expensive.

1.7.2 Iterative solvers

Alternatively, it is possible to invert the original KKT matrix (1.15) instead of its Schur complement. Since it is not positive definite, only iterative solvers such as the Minimized Residual (MinRes) or Generalized Minimized Residual (GMRes) method can be used. For their efficiency, good preconditioners would be necessary. Their construction, though, is not simple due to the saddle-point structure and indefiniteness. In particular, MinRes requires a positive definite symmetric preconditioner. In general, solving the whole Newton system with an iterative solver is considered a hard problem, due to the size of the problem, its ill-conditioning, and the structure of the matrix, see Saad [56] and Haber and Ascher [37].

The most efficient solver for the whole linear problem would probably be a multigrid solver, or an iterative method preconditioned by multigrid. Unfortunately, the finite element library used in this work does not have multigrid methods fully implemented yet.

In absence of a multigrid solver, two linear solvers have been used in the programs that implement the methods of this section. The first is MinRes (see Paige and Saunders [53]) with a diagonal scaling as preconditioning. Even though the preconditioning improved the performance significantly, the method often did not converge in a number of iterations less than the size of the full Newton matrix. This makes the method unsuitable for the problems we consider.

As a second alternative, we also tried GMRes (see Saad [56]), which allows for non-symmetric and even indefinite preconditioners. We used ILU or Vanka type preconditioners [66], or, if multi-processor machines are available, block variants thereof. While it is known that Vanka type methods are better smoothers than solvers, even ILU did not yield good performance of the solver, due to the high cost of constructing and applying the preconditioner. For larger problems, GMRes did not converge in a reasonable number of iterations, too.

As a last method, we tried to use the CG method on the normal equations,

$$H^2 \, \delta x = H f,$$

with H the global matrix in (1.15). Unfortunately, H^2 is so ill-conditioned that the CG method either took many iterations, or failed altogether.

Concluding this section, neither choice of iterative linear solvers produced satisfactory results.

1.7.3 Direct solvers

Instead of the iterative solvers above, we also used direct solvers for the Newton matrices. Due to memory considerations and the complexity of the task, only solvers that take sparsity into account can be used.

In our experiments, we have used the sparse direct solvers MA27 and MA47 from the Harwell Subroutine Library (see Duff and Reid [30, 31, 40]). They are specialized to symmetric indefinite systems of linear equations, and use a sparse variant of Gaussian elimination (MA27) or a multifrontal Gaussian elimination solver with 2×2 pivots similar to the Bunch-Parlett factorization (MA47). The choice between the two algorithms depends on a trade-off between memory consumption and computing time: MA47 is often significantly faster, but takes much more memory (up to a factor of five) than MA27 to compute the sparse decomposition.

Although requiring significantly more memory than iterative solvers, the main advantage of the direct solvers is that they never fail to find the solution of the linear subproblems; iterative solvers sometimes break down or take an excessive number of iterations, in which case the Newton algorithm may also break down due to an insufficient search direction.

The computing time required by direct solvers is less than or comparable to that of iterative solvers for the whole system for sizes up to at least 10^5 degrees of freedom.

1.7.4 Stopping criteria for the linear solvers

Unless we use a direct solver for the linear system (1.15), we do not solve each Newton step to very high accuracies, since Newton updates only approximate the step to the solution of the stationarity condition anyway. Such methods are usually termed *truncated* or *inexact Newton methods*, see Nocedal and Wright [51].

In practice, the inner solution is stopped once the linear residual in the l_2 norm has been reduced by a certain factor, say 10^3. Since the size of the linear systems grows due to mesh refinement as the outer nonlinear iterations proceed, reduction by a fixed factor amounts to increasing accuracy per degree of freedom in the Newton updates, eventually turning the truncated into an exact Newton method.

1.8 Theoretical considerations

It is not at all trivial to infer that the method proposed above works from a theoretical point of view. Beyond what is covered in this work, there are several theoretical questions that we would like to touch as they are needed to guarantee convergence to the solution of the original continuous problem 1.7. Since they are beyond the scope of this work, we only mention them, without giving answers.

Existence, uniqueness, and stability of solutions. These questions are discussed in a very general framework in Kravaris and Seinfeld [47], and in the book by Banks and Kunisch [13], where many results are proven without reference to concrete functionals or spaces. These results can then be checked for actual applications. However, results of this type usually require unduly high smoothness.

Probably the most general existence result for the problem treated in this chapter is given in Chavent et al. [26], where it is shown that there exist solutions on the rather weak assumptions that $u^* \in H^1, a^* \in \mathcal{A} \subset \{\chi \in L^\infty, 0 < a_0 \leq \chi \leq a_1\}$ if $m(\varphi) = \frac{1}{2}\|\nabla\varphi\|^2$. For stability of solutions, refer to Theorems 1.9 and 1.10.

Validity of the Lagrange principle. The question whether the state equation constraint allows an augmentation to a Lagrangian including both the minimization functional as well as the augmented state equation is discussed extensively in papers dealing with the *Augmented Lagrangian* formulation of the parameter estimation problem, see for example Ito and Kunisch [42, 41].

We quote here Theorem 2.1 of Ito and Kunisch [42], in which existence and uniqueness of a Lagrange multiplier is proven for a particular choice of functionals:

Theorem 1.22 (Ito and Kunisch). *Let* $x = \{u, a, \lambda\} \in H_0^1 \times H^2 \times H_0^1$ *and*

$$m(u-z) = \tfrac{1}{2}|u-z|^2_{H^1}, \qquad r(a) = \tfrac{1}{2}\|\nabla a\|^2 + \tfrac{1}{2}\|\nabla^2 a\|^2,$$
$$L(x) = m(u-z) + \beta r(a) + (a\nabla u, \nabla\lambda) - (f,\lambda).$$

Then there exists a unique Lagrange multiplier λ^* *such that the solution* $x^* = \{u^*, a^*, \lambda^*\}$ *of Problem 1.7 is characterized by the first order conditions given in Problem 1.8.*

The proof is given in Ito and Kunisch [42] for $d \equiv \dim\Omega = 2, 3$, for which the Sobolev inequality $\|v\|_{L^\infty} \leq C\|v\|_{H^2}$ holds. For $d = 1$, one can also apply the theorem for $r(a) = \frac{1}{2}\|\nabla a\|^2$.

In the cited paper, it is also shown that constraints of the form $a \geq a_0$ can be treated as well by adding a corresponding term $\langle\mu, a - a_0\rangle_{H^2}$ to the Lagrangian, with a Lagrange multiplier $\mu \in \mathcal{C}^+$, with

$$\mathcal{C} = \{w \in H^2 : w \geq 0\}, \qquad \mathcal{C}^+ = \{\mu \in H^2 : \langle\mu, w\rangle \leq 0 \ \forall w \in \mathcal{C}\}.$$

This multiplier is shown to exist and to be unique.

Convergence of continuous Newton steps. Rates of convergence can usually be stated in the form of a so-called *source condition*: if F is the operator mapping the parameter to the state space, i.e. in the present context $F(a) = (-\nabla\cdot(a\nabla))^{-1} f : \mathcal{A} \to V_g$ with fixed f,

$$F'(a)\delta a = -[-\nabla\cdot(a\nabla)]^{-1}[-\nabla\cdot(\delta a\nabla)][-\nabla\cdot(a\nabla)]^{-1} f$$

its derivative in direction δa, and $F'(a)^*$ the adjoint, and if the difference between initial estimate a_0 and exact solution a^* allows a representation

$$a^* - a_0 \in \text{range}\left((F'(a^*)^* F'(a^*))^\nu\right) \tag{1.19}$$

for some real number $\nu \geq 0$, then under certain additional conditions (see, for example, Deuflhard et al. [28], and Kaltenbacher [46]) the rate of convergence is, even in the noise free case with $\delta = 0$, only

$$\|a_n - a^*\| = \mathcal{O}\left(n^{-\nu}\right), \qquad \|u_n - u^*\| = \mathcal{O}\left(n^{-\nu-1/2}\right),$$

where n denotes the number of the Newton step.

If we neglect the possibility that we put a priori knowledge of potential non-smoothness into the initial iterate a_0, the source condition can be interpreted as follows: since $F'(a)$ mapping from the tangent space $\mathcal{A}'[a]$ of $\mathcal{A}$ to V_g has smoothing properties, the condition requires a^* to be smooth in order to obtain reasonable rates of convergence, i.e. ν significantly greater than zero. If such smoothness is missing, then the rate of convergence can be arbitrarily slow.

As an example, for one dimensional problems, an index $\nu = \frac{1}{2}$ already corresponds to $a^* - a_0 \in \{a \in H^3 \cap H_0^1 : \int_\Omega \Delta a / \left[\nabla(-\nabla\cdot(a^*\nabla))^{-1} f\right] = 0\}$ while for $\nu = \frac{1}{4}$ the requirements are loosened to H^2 instead of H^3, see Kaltenbacher [46]. In practice, such smoothness requirements are rarely met. In general, we are thus only able to guarantee qualitative convergence $\|a_n - a^*\| = o(1)$.

Existence, uniqueness, and stability of discretized Newton directions. For the Gauß-Newton modification, existence and uniqueness of discrete search directions is given by Lemma 1.21 under reasonable conditions on the functionals. However, this is not sufficient in general, as we want a *stable* solution as the mesh width $h \to 0$. For this case, refer to Banks and Kunisch [13].

Convergence of discrete solutions. As we generate a sequence of solutions a_h^* on successively refined meshes, we are interested in rates of convergence against the solution a^* of the continuous problem. Such rates are proven in Falk [34], and also in Banks and Kunisch [13, Theorem IV.3.1 and Remark IV.3.6], but rely on rather strong assumptions on the smoothness of the unknown solution a^* and the proofs in Banks and Kunisch [13] also require to use H^2 finite elements. For completeness, we briefly restate Theorem IV.3.1 from [13], which is for the case of no regularization:

Theorem 1.23 (Banks and Kunisch). *Let $s > 1$. If $a^* \in H^s$, $f \in H^{s-1}$, $z \in H^{s+1}$, $u^* \in H^{s+1} \cap W^{2,p}, p > \dim\Omega$, where $u^* = [-\nabla\cdot(a^*\nabla)]^{-1} f$. Let a_h, u_h be discretized by finite elements with quadratic convergence order in the L^2 norm, and under additional smoothness assumptions on the finite element spaces, there exists $C > 0$ such that the following weighted estimate holds:*

$$\left\|(a_h^* - a^*)|\nabla u^*|^2\right\|_{L^1(\Omega)} < C\left(h^{-2}d + h^{s-1}\right),$$

where d is the distance of the measurement z from the attainable set $\{u : -\nabla\cdot(a\nabla u) = f, \text{for } a \in \mathcal{A}\}$.

Obviously, these smoothness requirements are too strong for practical purposes. For actual smoothness levels, rates of convergence are coupled to the

index of the source condition discussed above. In most cases, only qualitative convergence can be expected.

For numerical indications to this phenomenon, see Section 2.1.3 and in particular Fig. 2.2, where we display the convergence of the parameter in computations for the test cases defined in Section 1.9; only for the first test case and in the noise free case is the source condition satisfied, with $\nu = \frac{1}{2}$. It is not satisfied for all other test cases even in the noise free case since there $a^* \notin H^1$.

Note also that the theorem states that without regularization we have to expect a deterioration of approximation under mesh refinement if the measurement is not attainable; this, as well, coincides with practical experience. For the reason why the weighting in the norm of the estimate is necessary, see Section 4.5.

Convergence of discretized Newton steps. As we do not solve the exact Newton step (1.12) but a discrete approximation of it, we have to show that Newton's method still converges to the correct solution (at least if $h \to 0$ as we proceed with Newton steps). In finite dimensional optimization, it is usually shown that the true and the approximate KKT matrix do not differ too much, i.e. here

$$\|\nabla L(x_k) - \tilde{H}_k(x_k - x^*)\| \leq C\|x_k - x^*\|^2.$$

where $\tilde{H}_k = \left[(P_h \nabla_x L(x_k))^\dagger P_h\right]^{-1}$ is the discretized Hessian, P_h is the $\mathcal{X}_g$-orthogonal projector onto the finite dimensional subspace $\mathcal{X}_h$, and $B^\dagger$ is the generalized inverse of B.

While this condition is difficult to prove for the present context, it is also not very appropriate in the context of ill-posed problems. For a discussion of this topic, see Kaltenbacher [46, Section 2.1].

1.9 Definition of test cases

In the following chapters, we will demonstrate various aspects of the methods discussed at some test cases, which we define in this section. Parameters and state variables are plotted in Fig. 1.2 for the different test cases and for $\mathbf{x} \in \mathbb{R}^2$.

Test case 1.1 (Smooth parameter). *Let*

$$a(\mathbf{x}) = 1 + |\mathbf{x}|^2, \qquad u(\mathbf{x}) = |\mathbf{x}|^2, \qquad f = -\nabla \cdot (a \nabla u).$$

On the boundary $\Gamma_D = \partial\Omega$, we set $g = u$.

Test case 1.2 (Discontinuous parameter). *Let*

$$a(\mathbf{x}) = \begin{cases} 1 & \text{for } |\mathbf{x}| < \frac{1}{2} \\ 8 & \text{else,} \end{cases} \qquad u(\mathbf{x}) = \begin{cases} |\mathbf{x}|^2 & \text{for } |\mathbf{x}| < \frac{1}{2} \\ \frac{1}{8}|\mathbf{x}|^2 + \frac{7}{32} & \text{else,} \end{cases}$$

and $f = -\nabla \cdot (a \nabla u) = -2d$ for $\mathbf{x} \in \mathbb{R}^d$. Note that here the locations of discontinuities in a and in ∇u match, and the right hand side is a smooth function; this matches the case usually found in stationary physical applications. We choose as Dirichlet boundary $\Gamma_D = \partial\Omega$, with $g = u$ there.

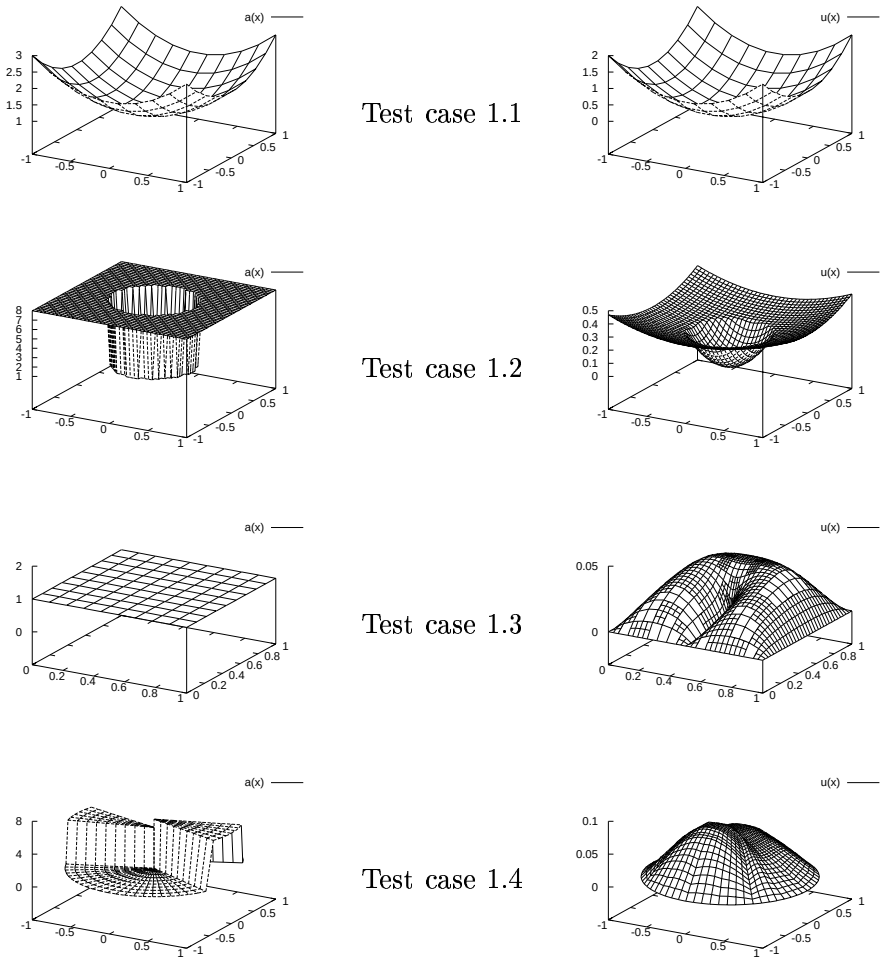


Figure 1.2: *Parameter* $a(\mathbf{x})$ *(left) and state variable* $u(\mathbf{x})$ *(right) for the different test cases.*

Test case 1.3 (Singular solution). *Let* Ω *be the slit domain* $(0,1)^d \backslash \{x = \frac{1}{2}, y \leq \frac{1}{2}\}$, *and*

$$a = 1, \qquad f = 1, \qquad u = [-\nabla\cdot(a\nabla)]^{-1} f, \qquad u|_{\partial\Omega} = 0.$$

For this example, the quantitative resolution of the singularity is decisive for efficient algorithms. Although the coefficient is constant, we discretize it as a distributed one as for the other test cases.

Test case 1.4 (Criss-cross parameter). *Let* $\Omega = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| < 1\}$, *and*

$$a \in \{1, 2, 6, 7\}, \qquad f = 1 - \|\mathbf{x}\|^2, \qquad u = [-\nabla\cdot(a\nabla)]^{-1} f, \qquad u|_{\partial\Omega} = 0.$$

The coefficient has piecewise constant values in the four sectors of the domain divided by the lines $y = \pm\frac{x}{3}$, as shown in Fig. 1.2. For this case, a singularity in u is generated at the point where different values of the coefficient meet.

For all test cases, the measurement z is obtained from the exact displacement u by adding some noise:

$$z(\mathbf{x}) = u(\mathbf{x}) + \delta(\mathbf{x}),$$

The noise $\delta(\mathbf{x})$ is a Gaussian random function with zero mean.

We remark that even in the noise free case, i.e. $\delta = 0$, the optimal solution $\{u^*, a^*\}$ of Problem 1.7 is not identical to the functions $\{u, a\}$ defined above if we add regularization, i.e. $\beta \neq 0$.

Chapter 2

Error estimates and adaptivity

In this chapter, we discuss error estimates and strategies for refinement of the discretization. We will primarily base these strategies on error representations derived using duality arguments, see Becker and Rannacher [17] and Becker [15], but will also consider other approaches such as stability or Lagrange multiplier estimates corresponding to discretization constraints.

Starting this chapter, we discuss error representation formulae with respect to the minimization functional $J(\cdot)$, and corresponding mesh refinement criteria. For this particular case, the use of weighted error estimates does not involve the solution of an additional problem when solving optimization problems. Thus, the evaluation of the error estimates basically comes at the same price as the evaluation of heuristic indicators. The resulting indicator is compared to other indicators with respect to its efficiency, and its reliability as an error estimator is verified.

After this, we derive estimates and criteria for the coefficient parameterization. As the discretization of the parameter variable is chosen mostly independent of that of the state variable, criteria for this particular purpose may be best suited for this. Again, we compare estimates and accuracy for efficiency.

We then consider estimates based on stability and estimates based on techniques involving the dual problem to the first order necessary conditions. These allow for error representation formulae and refinement criteria tailored to arbitrary functionals of the solution. Finally, estimates for the problem with constraints on the parameter are discussed.

To the author's best knowledge, there is nothing in the available literature where adaptive methods based on the actual optimization problem are employed for distributed parameter estimation problems, despite their obvious applicability in many cases. There are, however, some uses for optimization problems, see for example Becker et al. [16, 17, 15].

2.1 Error estimates for the minimization functional

In this section, we will derive a representation of the error in the minimization functional J defined in Problem 1.7, i.e. for the quantity $J(x) - J(x_h)$, where x and x_h are continuous and discrete solutions, respectively. First, we state

its abstract form only involving the Lagrangian of the problem (Theorem 2.1), then specialize it for the elliptic problem introduced in the previous chapter (Theorem 2.2). We will then discuss two ways for the practical evaluation of this error representation, assess their practical performance compared to more heuristic approaches, and also check their efficiency as error estimates.

2.1.1 Derivation of estimates

For the derivation of an error representation formula, recall that continuous and discrete solutions satisfy the variational equalities

$$\nabla_x L(x;y) = 0 \qquad \forall y \in \mathcal{X}_0, \tag{2.1}$$

$$\nabla_x L(x_h;y_h) = 0 \qquad \forall y_h \in \mathcal{X}_h, \tag{2.2}$$

respectively. The definition of the Lagrangian and of the function spaces is given in Problem 1.8. With these equalities, Galerkin orthogonality for this nonlinear problem reads:

$$\nabla_x L(x;y_h) - \nabla_x L(x_h;y_h) = 0 \qquad \forall y_h \in \mathcal{X}_h. \tag{2.3}$$

Using this identity, an expression for the error in the target functional is derived in the following theorem.

Theorem 2.1. *Let x and x_h be solutions to (2.1) and (2.2), respectively. Then the discretization error with respect to J is given with $e = x - x_h$ by*

$$J(x) - J(x_h) = \tfrac{1}{2}\nabla_x L(x_h; x - y_h) + R(x, x_h) \qquad \forall y_h \in \mathcal{X}_h, \tag{2.4}$$

where the remainder term $R(x, x_h)$ is given by

$$R(x, x_h) = \tfrac{1}{2}\int_0^1 \nabla_x^3 L(x_h + se; e, e, e)\; s(s-1)\; ds.$$

Proof. At the solution points the state equations are satisfied, therefore

$$J(x) - J(x_h) = L(x) - L(x_h).$$

On the other hand,

$$L(x) - L(x_h) = \int_0^1 \nabla_x L(x + se; e) ds,$$

with $e = x - x_h$, and by approximation by the trapezoidal rule

$$L(x) - L(x_h) = \tfrac{1}{2}\nabla_x L(x; e) + \tfrac{1}{2}\nabla_x L(x_h; e) + \tfrac{1}{2}\int_0^1 \nabla_x^3 L(x_h + se; e, e, e)\; s(s-1)\; ds.$$

The first term vanishes by the optimality condition (2.1). In view of Galerkin orthogonality (2.3) and the discrete identity (2.2), we have that

$$\begin{aligned}\nabla_x L(x_h; e) &= \nabla_x L(x_h; x) - \nabla_x L(x_h; x_h) = \nabla_x L(x_h; x) \\ &= \nabla_x L(x_h; x) - \nabla_x L(x_h; y_h) = \nabla_x L(x_h; x - y_y)\end{aligned}$$

for any $y_h \in \mathcal{X}_h$. The assertion then follows. □

For the diffusion equation introduced in the previous chapter, and for the particular case that misfit and regularization functionals are quadratic, the error representation (2.4) with an arbitrary $y_h = \{\varphi_h, \chi_h, \psi_h\}$ has the following form:

$$J(x) - J(x_h) = \frac{1}{2}\Big\{\rho_u(x_h; x - y_h) + \rho_\lambda(x_h; x - y_h) + \rho_a(x_h; x - y_h)\Big\} + R \tag{2.5}$$

with residuals

$$\begin{aligned}
\rho_u(x_h; x - y_h) &= m'(u_h - z; u - \varphi_h) + (a_h \nabla\lambda_h, \nabla(u - \varphi_h)),\\
\rho_\lambda(x_h; x - y_h) &= (a_h \nabla u_h, \nabla(\lambda - \psi_h)) - (f, \lambda - \psi_h),\\
\rho_a(x_h; x - y_h) &= \beta r'(a_h; a - \chi_h) + (\nabla\lambda_h \cdot \nabla u_h, a - \chi_h)),
\end{aligned}$$

and remainder term

$$R = -\frac{1}{12}((a - a_h)\nabla(\lambda - \lambda_h), \nabla(u - u_h)).$$

The remainder term does not contain intermediate points any more, since the state equation was assumed to be quadratic.

From this representation, we can obtain a *localized* error estimate. We demonstrate this for a particular choice of discretization spaces and functionals, but it is straightforward to generalize it to other situations.

Theorem 2.2. *Let misfit and regularization functional be*

$$m(u - z) = \tfrac{1}{2}\|u - z\|^2, \qquad r(a) = \tfrac{1}{2}\|a\|^2.$$

Then, the following error representation holds:

$$\begin{aligned}
J(x) - J(x_h) = \frac{1}{2}\sum_{K\in\mathbb{T}} \Big\{ & (-f - \nabla\cdot(a_h\nabla u_h), \lambda - i_h\lambda)_K + \tfrac{1}{2}\,(\mathbf{n}\cdot[a_h\nabla u_h], \lambda - i_h\lambda)_{\partial K}\\
& + (u_h - z - \nabla\cdot(a_h\nabla\lambda_h), u - i_h u)_K + \tfrac{1}{2}\,(\mathbf{n}\cdot[a_h\nabla\lambda_h], u - i_h u)_{\partial K}\Big\}\\
& + \frac{1}{2}\sum_{K_a\in\mathbb{T}_a} (\beta a_h + \nabla\lambda_h\cdot\nabla u_h, a - i_h a)_{K_a}\\
& - \frac{1}{12}((a - a_h)\nabla(\lambda - \lambda_h), \nabla(u - u_h)),
\end{aligned} \tag{2.6}$$

with a generic interpolation operator i_h acting on $\mathcal{X} \to \mathcal{X}_h$ or single components, depending on context. For edges $\gamma \subset \partial K$ between a cell K and a neighbor K', we define the jump terms by

$$\mathbf{n}\cdot[a_h\nabla\varphi_h] = \begin{cases} \mathbf{n}\cdot(a_h|_{K'}\nabla\varphi_h|_{K'} - a_h|_K\nabla\varphi_h|_K) & \text{if } \gamma \not\subset \partial\Omega,\\ 2\mathbf{n}\cdot a_h\nabla\varphi_h & \text{if } \gamma \subset \partial\Omega. \end{cases}$$

Proof. Split the integrals in (2.5) into sums over all cells and integrate by parts on each cell. Then exchange half of the boundary terms on each cell with the neighbors to obtain optimal order locally. Set $y_h = i_h x$. □

Since the error representation above involves the exact solution x, we evaluate it approximatively by using a guess $\tilde{x}$ of x; for techniques to obtain such guesses, we refer to the overview article by Becker and Rannacher [17]. With this, we define the following *approximate error representation* by replacing x by $\tilde{x} = \{\tilde{u}, \tilde{a}, \tilde{\lambda}\}$, $i_h x$ by x_h, and neglecting the remainder term:

$$\begin{aligned}
\eta^{DWR1} &= \sum_{K\in\mathbb{T}} \eta_K + \eta_{\partial K} + \sum_{K_a\in\mathbb{T}_a} \eta_{K_a},\\
\eta_K &= \frac{1}{2}\left\{ (u_h - z - \nabla\cdot(a_h\nabla\lambda_h), \tilde{u} - u_h)_K - \left(f + \nabla\cdot(a_h\nabla u_h), \tilde{\lambda} - \lambda_h\right)_K \right\},\\
\eta_{\partial K} &= \frac{1}{2}\left\{ \tfrac{1}{2}\left(\mathbf{n}\cdot[a_h\nabla\lambda_h], \tilde{u} - u_h\right)_{\partial K} + \tfrac{1}{2}\left(\mathbf{n}\cdot[a_h\nabla u_h], \tilde{\lambda} - \lambda_h\right)_{\partial K} \right\},\\
\eta_{K_a} &= \frac{1}{2}\sum_{K_a\in\mathbb{T}_a} \left(\beta a_h + \nabla\lambda_h\cdot\nabla u_h, \tilde{a} - a_h\right)_{K_a}.
\end{aligned} \tag{2.7}$$

If we cannot or do not want to provide a guess $\tilde{x}$ for x, then the following theorem may still help us to develop an error *estimate*:

Theorem 2.3. *Let $\mathcal{U}_h = \Lambda_h = Q_c^1(\mathbb{T})$, $\mathcal{A}_h = Q_d^0(\mathbb{T}_a)$. Under the same assumptions as in Theorem 2.2, and assuming that for the exact solution we have $u, \lambda \in H^2$, $a \in H^1$, there holds the following a posteriori estimate for the error:*

$$\begin{aligned}
|J(x) - J(x_h)| &\le \eta + \frac{1}{12}|((a - a_h)\nabla(\lambda - \lambda_h), \nabla(u - u_h))|\\
\eta &= C_I^1 \sum_{K\in\mathbb{T}} \left(\rho_K^u\omega_K^u + \rho_{\partial K}^u\omega_{\partial K}^u + \rho_K^\lambda\omega_K^\lambda + \rho_{\partial K}^\lambda\omega_{\partial K}^\lambda\right) + C_I^2 \sum_{K_a\in\mathbb{T}_a} \rho_{K_a}^a\omega_{K_a}^a,
\end{aligned} \tag{2.8}$$

with residuals and weights

$$\begin{aligned}
\rho_K^u &= \tfrac{1}{2}\|u_h - z - \nabla\cdot(a_h\nabla\lambda_h)\|_K, & \omega_K^u &= h_K^2\left\|\nabla^2 u\right\|_K,\\
\rho_{\partial K}^u &= \frac{1}{4}\|\mathbf{n}\cdot[a_h\nabla\lambda_h]\|_{\partial K}, & \omega_{\partial K}^u &= h_K^{3/2}\left\|\nabla^2 u\right\|_K,\\
\rho_K^\lambda &= \tfrac{1}{2}\|f + \nabla\cdot(a_h\nabla u_h)\|_K, & \omega_K^\lambda &= h_K^2\left\|\nabla^2\lambda\right\|_K,\\
\rho_{\partial K}^\lambda &= \frac{1}{4}\|\mathbf{n}\cdot[a_h\nabla u_h]\|_{\partial K\setminus\partial\Omega}, & \omega_{\partial K}^\lambda &= h_K^{3/2}\left\|\nabla^2\lambda\right\|_K,\\
\rho_{K_a}^a &= \tfrac{1}{2}\|\beta a_h + \nabla\lambda_h\cdot\nabla u_h\|_{K_a}, & \omega_{K_a}^a &= h_{K_a}\|\nabla a\|_{K_a}.
\end{aligned}$$

From a practical point of view, the interpolation constants C_I^1, C_I^2 are usually in the range $0.1 \ldots 1$.

Proof. Use the Cauchy-Schwartz inequality to separate the scalar products in (2.6). Assuming the indicated regularity of the exact solution, we can use the Bramble-Hilbert lemma to estimate $\|u - i_h u\|_K \le C h_K^2 \|\nabla^2 u\|_K$, $\|u - i_h u\|_{\partial K} \le C h_K^{3/2}\|\nabla^2 u\|_K$, and likewise for λ, and $\|a - i_h a\|_{K_a} \le C h_{K_a}\|\nabla a\|_{K_a}$, where i_h is a generic interpolation operator $V \to V_h$ or $\mathcal{A} \to \mathcal{A}_h$, depending on its argument.

For $\rho^\lambda_{\partial K}$, note that for faces $\partial K \subset \partial\Omega$ there holds $\lambda - i_h\lambda|_{\partial K} = 0$ since $\lambda|_{\partial\Omega} \equiv 0$. Thus, these jump residuals give no contribution at the boundary, which we take into account by setting them to zero since this information is lost when estimating $\|\lambda - i_h\lambda\|_{\partial K} = 0$ by $Ch_K^{3/2}\|\nabla^2\lambda\|_K \geq 0$. Note, however, that this does not hold for $\rho^u_{\partial K}$ since in general $u - i_h u \neq 0$ at $\partial\Omega$. □

Again, the weights ω contain the exact solution x. However, since no relation to the discrete space $\mathcal{X}_h$ is involved this time, we can hope to get a good approximation of η by substituting $\|\nabla^2 u\|_K \to \|\nabla_h^2 u_h\|_K$ with some discrete approximation ∇_h to the gradient ∇, e.g. a difference quotient, and likewise for the norms in the other weights. For reference below, we define the following *approximate error estimate* using this substitution:

$$\eta^{DWR2} = C_I^1 \sum_{K\in\mathbb{T}} \left(\rho^u_K\tilde\omega^u_K + \rho^u_{\partial K}\tilde\omega^u_{\partial K} + \rho^\lambda_K\tilde\omega^\lambda_K + \rho^\lambda_{\partial K}\tilde\omega^\lambda_{\partial K}\right) + C_I^2 \sum_{K_a\in\mathbb{T}_a} \rho^a_{K_a}\tilde\omega^a_{K_a}, \tag{2.9}$$

with residuals and approximate weights defined by

$$\begin{aligned}
\rho^u_K &= \tfrac{1}{2}\|u_h - z - \nabla\cdot(a_h\nabla\lambda_h)\|_K, & \tilde\omega^u_K &= h_K^2\left\|\nabla_h^2 u_h\right\|_K,\\
\rho^u_{\partial K} &= \frac{1}{4}\|\mathbf{n}\cdot[a_h\nabla\lambda_h]\|_{\partial K}, & \tilde\omega^u_{\partial K} &= h_K^{3/2}\left\|\nabla_h^2 u_h\right\|_K,\\
\rho^\lambda_K &= \tfrac{1}{2}\|f + \nabla\cdot(a_h\nabla u_h)\|_K, & \tilde\omega^\lambda_K &= h_K^2\left\|\nabla_h^2 \lambda_h\right\|_K,\\
\rho^\lambda_{\partial K} &= \frac{1}{4}\|\mathbf{n}\cdot[a_h\nabla u_h]\|_{\partial K\setminus\partial\Omega}, & \tilde\omega^\lambda_{\partial K} &= h_K^{3/2}\left\|\nabla_h^2 \lambda_h\right\|_K,\\
\rho^a_{K_a} &= \tfrac{1}{2}\|\beta a_h + \nabla\lambda_h\cdot\nabla u_h\|_{K_a}, & \tilde\omega^a_{K_a} &= h_{K_a}\|\nabla_h a_h\|_{K_a}.
\end{aligned}$$

Remark 2.4. *The regularity assumed in Theorem 2.3 is not very practical. In particular, since the Lagrange multiplier has to satisfy the equation*

$$-\nabla\cdot(a\nabla\lambda) = -(u - z),$$

it will not be in H^2 if the optimal coefficient a is not smooth, or if the domain Ω is not convex. Similar considerations hold for u. Nevertheless, taking difference quotients in the weights in (2.9) *is well-defined and yields at places of missing regularity negative powers of the mesh width, resulting locally in the correct order.*

2.1.2 Criteria for refinement of the state mesh

In this section, we propose several refinement criteria for the state equation mesh $\mathbb{T}$. We will then compare these for example problems.

Refinement indicator η_K^{DWR1} (dual weighted residuals). Starting from the representation (2.6), we use the approximate error representation (2.7) as refinement indicator.

Refinement indicator η_K^{DWR2} (dual weighted residuals). We can also base a refinement criterion on the error estimate (2.8), and use the approximate error estimate (2.9) as refinement indicator. Since the meshes produced by this and the previous refinement criterion perform almost identical, we do not list this indicator in most charts.

Refinement indicator $\eta_K^{\nabla\nabla u}$ (smoothness of u_h). As a first heuristic refinement criterion, we may use an indicator measuring solely the smoothness of the primal variable:

$$\eta_K^{\nabla\nabla u} = h_K^{(d+3)/2} \|\nabla_h^2 u_h\|_K. \tag{2.10}$$

This indicator is well known from the Laplace equation.

Besides the heuristic argument, the indicator can be made plausible by simplification of the dual error estimator (2.9): assume $\lambda \in H^2$ and $\nabla\cdot(a\nabla u) \in L^\infty(\Omega)$, and assume convergence of the term $\|\mathbf{n}\cdot[a_h\nabla\lambda_h]\|_{\partial K} \to \|h\nabla\cdot(a\nabla\lambda)\|_K \leq h^{d/2}\|\nabla\cdot(a\nabla\lambda)\|_{\infty;K} = h^{d/2}\|u-z\|_{\infty;K} \leq h^{d/2}\|u-z\|_{\infty;\Omega} \leq c_s h^{d/2}$ with a stability constant $c_s = \|u-z\|_{\infty;\Omega}$. Then the second term in the error bound (2.8) can be estimated as

$$\rho_{\partial K}^u \omega_{\partial K}^u \leq c_s h_{\tilde{K}}^{(d+3)/2} \|\nabla^2 u\|_{\tilde{K}}.$$

The indicator (2.10) then arises by using finite difference quotients instead of derivatives, replacing the exact value u by u_h, and dropping the constant factor c_s which is irrelevant for refinement.

We would like to stress that the derivation sketched above is rather heuristic and does not stand formal criteria. For example, numerical experiments suggest that in general, the assumed convergence $\|\mathbf{n}\cdot[a_h\nabla\lambda_h]\|_{\partial K} \to \|h\nabla\cdot(a\nabla\lambda)\|_{\partial K}$ does not hold on non-uniform, possibly locally refined meshes with hanging nodes. However, refinement indicators like the one shown above are used successfully in practice. Therefore, we use them for comparison.

Refinement indicator $\eta_K^{\nabla\nabla\lambda}$ (smoothness of λ). Using a similar line of reasoning, take the first term in (2.8) and obtain the following refinement indicator:

$$\eta_K^{\nabla\nabla\lambda} = h_K^{(d+3)/2} \|\nabla_h^2 \lambda_h\|_K, \tag{2.11}$$

2.1.3 Comparison of refinement criteria

The performance of the various refinement criteria with respect to the reduction of $J(x_h)$ and the resolution of the unknown parameter is compared in Fig.s 2.1 and 2.2, using the test cases defined in Section 1.9.

Before discussing the results, we note that driving refinement by setting up an error estimator for the value of $J(\cdot)$ is, beyond the fact that it is essentially for free, reasonable since the value of $J(x_h)$ may be used to stop an iteration if it falls below the noise level. As $m(u-z)$ is bounded from below by noise, we would only resolve this noise if we reduced J further. However, this would

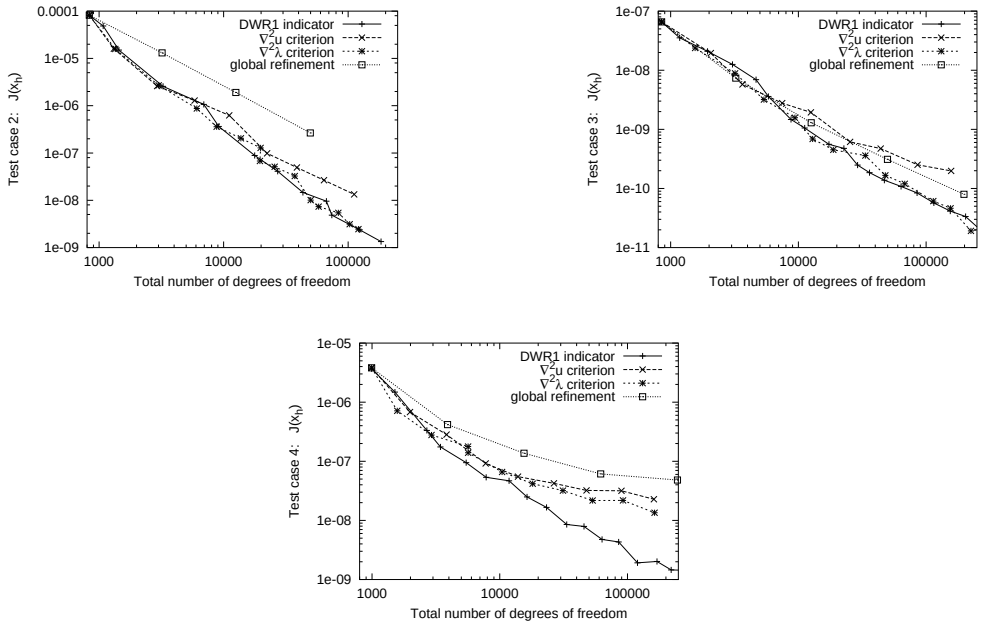


Figure 2.1: *Comparison of value of the minimization functional $J(x_h)$ for various refinement criteria. Top left: Test case 2 (discontinuous coefficient). Top right: Test case 3 (slit domain). Bottom: Test case 4 (criss-cross parameter).*

not lead to a better resolution of the parameter. Monitoring the value $J(x_h)$ and comparing it with an improved estimate therefore helps to stop iterations when this happens.

The results of computations are visualized in Fig.s 2.1 and 2.2. They can be summarized as follows:

- The criterion η_K^{DWR1} based on the dual error representation formula performs better than or equal to all other criteria under investigation for all examples.
- For most examples, the dual weighted error estimate and the $\eta_K^{\nabla\nabla\lambda}$ indicator perform equally well. They are always better than the other refinement indicators.
- Only for test case 4 is the dual weighted estimate significantly better than $\eta_K^{\nabla\nabla\lambda}$.

Meshes generated by the various refinement criteria are shown in Fig.s 2.3 and 2.4 for test cases 3 and 4. They are only slightly different for all test cases, even for test case 4 where the duality based estimator is significantly better quantitatively.

The fact that the dual weighted error estimate does not perform better as mesh refinement criterion than the more ad hoc indicator $\eta_K^{\nabla\nabla\lambda}$, defeats intuition at first. However, comparing the relative sizes of the contributions to the dual weighted error representation (2.7) reveals that in actual computations the

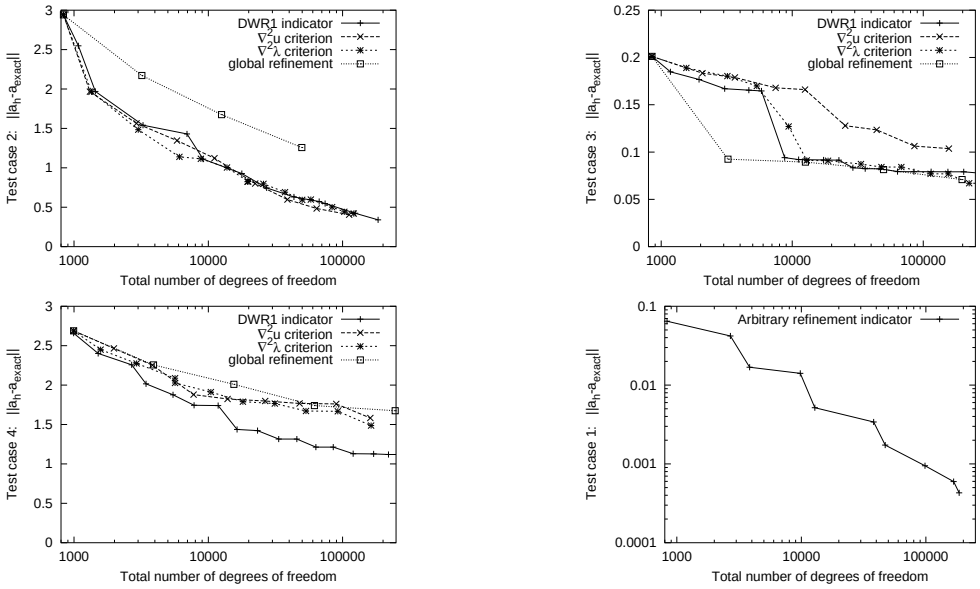


Figure 2.2: *Comparison of various mesh refinement criteria with respect to the error in the coefficient* $\|a_h - a_{exact}\|$. *Top row: Test cases 2 and 3. Bottom left: Test case 4. Bottom right: For comparison* $\|a_h - a_{exact}\|$ *for test case 1, where all refinement indicators work equally well. Note that here the error is scaled logarithmically (see also the discussion of condition* (1.19)*).*

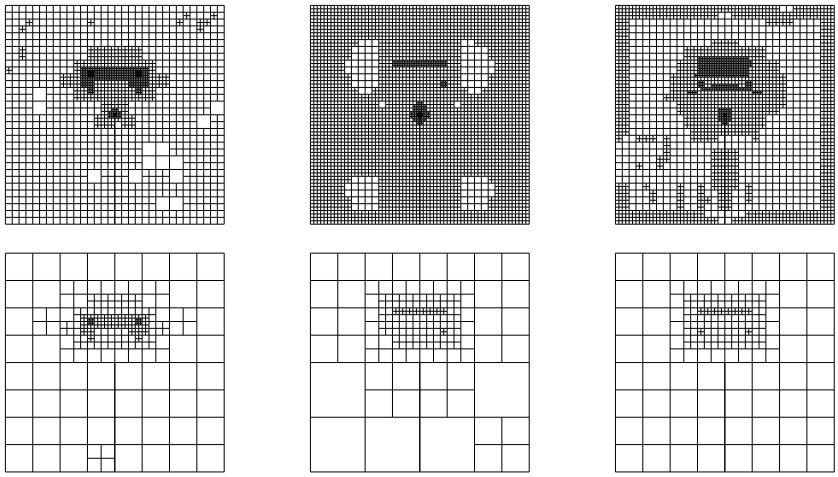


Figure 2.3: Test case 3 (slit domain): Comparison of meshes generated by criteria η_K^{DWR1}, $\eta_K^{\nabla\nabla u}$, $\eta_K^{\nabla\nabla\lambda}$ (from left to right). Top row: Meshes $\mathbb{T}$ for state and adjoint variable. Bottom row: Meshes $\mathbb{T}_a$ for the parameter a.

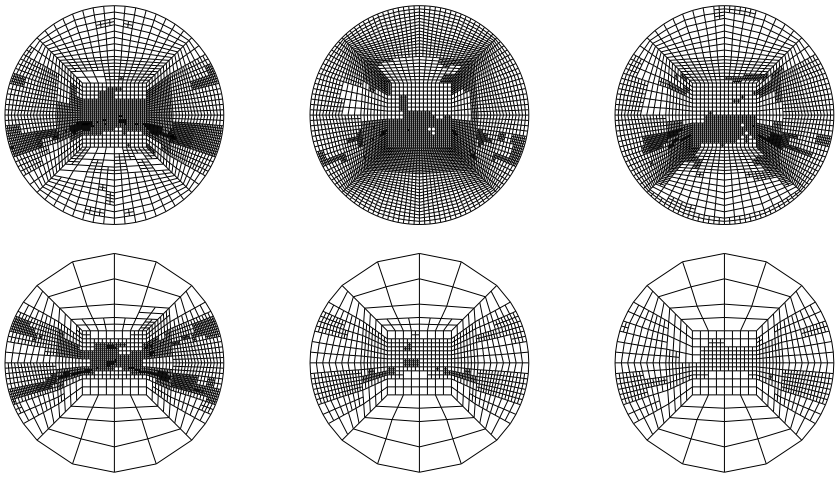


Figure 2.4: Test case 4 (crisscross parameter): Comparison of meshes generated by criteria η_K^{DWR1}, $\eta_K^{\nabla\nabla u}$, $\eta_K^{\nabla\nabla\lambda}$ (from left to right). Top row: Meshes $\mathbb{T}$ for state and adjoint variable. Bottom row: Corresponding meshes $\mathbb{T}_a$ for the parameter a.

term η_{K_a} is small compared to the other terms, at least in those cases where the two refinement criteria perform equally. On the other hand, second derivatives of the Lagrange multiplier or comparable terms appear in the two other terms $\nabla_u L(x_h; u - i_h u)$ and $\nabla_\lambda L(x_h; \lambda - i_h \lambda)$, either as residuals or weights. Since these terms in the weighted estimator consist of products of functions of u and of λ, it can only show fundamentally different behavior than the $\eta_K^{\nabla\nabla\lambda}$ indicator if the regions of roughness of u and λ do not coincide. However, this can not happen since the Lagrange multiplier satisfies $-\nabla\cdot(a\nabla\lambda) = -(u - z)$, and if no noise is present then $u - z$ is proportional to $\nabla^2 u$, i.e. u and λ have the same local smoothness properties.

On the other hand, for test case 4, where the dual weighted estimator performed better, the term η_{K_a} in (2.7) is *not* small compared to the other terms. These considerations explain why the dual weighted indicator and the $\eta_K^{\nabla\nabla\lambda}$ indicator perform equally well in most situations, and in which situations the former is better.

2.1.4 Reliability of error estimates

Besides providing refinement criteria, the error indicators (2.7) and (2.9) may be used to assess the quality of the finite element approximation x_h of (2.2) with respect to the true solution $x = \{u, a, \lambda\}$ of (2.1). In this section, we discuss how reliable these estimates for the quantity $J(x) - J(x_h)$ are.

Since for the general problem the exact solution is usually unknown, we restrict ourselves to the case of $\beta = 0$, and that z is a feasible point. We can then assume that we can find a parameter a such that for the corresponding primal variable $u = z$ holds, and thus $m(u - z) = 0$. Since $\beta = 0$ we have that $J(x) = 0$ and the exact error is given by $J(x) - J(x_h) = -J(x_h) = -m(u_h - z)$.

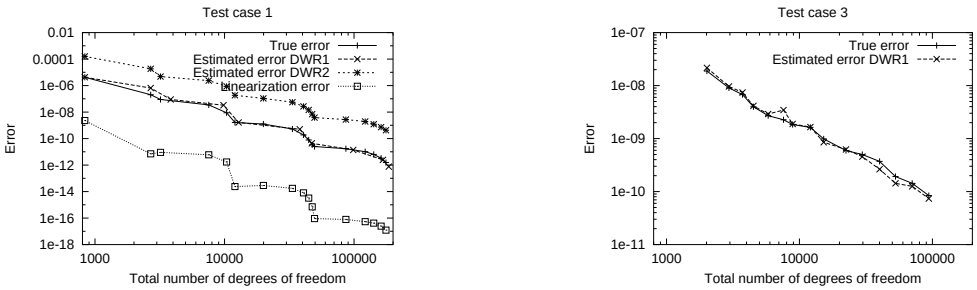


Figure 2.5: *Comparison of error estimates DWR1* (2.7) *and, for test case 1, DWR2* (2.9) *with the approximate true error* $\tilde{\mathcal{E}}$. *For the first example we also show the linearization error in* (2.5).

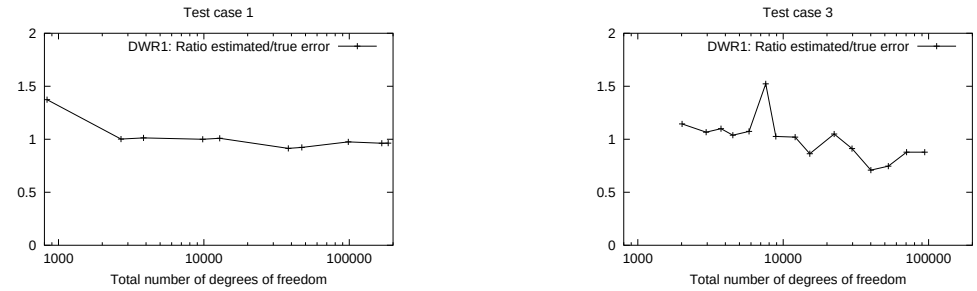


Figure 2.6: *Ratio of error estimates* (2.7) *(DWR1) and the approximate true error* $\tilde{\mathcal{E}}$, *for test cases 1 and 3.*

Assuming that the scheme converges to the global solution, we can then compare the error estimates with this value.

On the other hand, if $\beta > 0$, then at the solution $\beta r(a) > 0$, and $m(u-z) > 0$ since in general $u \neq z$. The exact error is then unknown. However, if β is small, the noise level large, or the computational mesh coarse, then $m(u_h-z) \gg \beta r(a_h)$ in the range of the x_h which we resolve in the course of our computations, and we can still expect that the quantity $\tilde{\mathcal{E}} = -m(u_h - z)$ is a good approximation to the true error $\mathcal{E} = J(x) - J(x_h)$. In Fig.s 2.5 and 2.6 we compare this value $\tilde{\mathcal{E}}$ with the estimates (2.7) and (2.9).

It is seen that the error estimates using (2.7) are in very good agreement with the actual error for test cases 1 and 3, showing the same convergence behavior and having a ratio between estimated and true error very close to the optimal value of 1. For test cases 2 and 4, where bounds are posed on the unknown solution, the estimates are unreliable; an extension of the estimates for the constrained problem is discussed in Section 2.5.

For test case 1, Fig. 2.5 also shows the values of estimate (2.9) where we have taken residuals and weights apart by the Cauchy-Schwarz inequality. We have chosen the interpolation constants equal to $C_I = 0.3$. As seen from the figure, the estimates are too large, with overestimation factors growing from 50

to roughly 250 under mesh refinement. For other examples, the ratio usually remains bounded, but is significantly too large as well.

Finally, Fig. 2.5 shows that the linearization term in (2.5), here computed using the exact solution, is sufficiently small that neglecting it in the error estimates is justified. This also holds for the other test cases.

2.2 Error estimates and adaptivity for the coefficient parameterization

In this section, we will describe methods of refinement of the parameter mesh $\mathbb{T}_a$. We will discuss an idea to use linearized sensitivities to refine the mesh based on a novel approach considering *discretization* as a constraint. Alternatively, mesh refinement will be based on heuristic arguments, or, if available, on information from the dual weighted residual error estimator derived in the last section.

Intuitively, one would like to base mesh refinement for the parameterization on sensitivities with respect to the state equation: we should refine the mesh where we know that the parameters are resolved best. For the discrete problem, the uncertainties are computed from the diagonal elements of the covariance matrix

$$C_{M'} = (C^T A^{-T} M A^{-1} C)^{-1},$$

see Tarantola [63]. Thus, the covariance matrix is the inverse of part of the Schur complement of the Gauß-Newton matrix which we need in each step anyway. Given the complexity of computing $C_{M'}$ (this would involve n forward and n backward solutions), this approach is not feasible, though. A second drawback is that it is not clear that refining where sensitivities are high is also necessarily a good strategy for the approximation of the parameter. For these reasons, we have used alternative refinement criteria for the parameter mesh, which we will discuss below.

2.2.1 Criteria based on discretization constraints

Here, we will first derive refinement criteria based on an unconventional approach in which we consider sensitivities with respect to discretization, which we take as a constraint here. It will be shown that refinement indicators can be based on the Lagrange multipliers associated with the discretization constraint. We show the derivation of such criteria for the unconstrained case, show an a posteriori bound on the error, and then extend the method to the constrained case.

Since we are only concerned with the parameter discretization, assume for the derivation that the parameter is discretized, while state and adjoint variable may or may not be discretized but that the space $\tilde{V} = V$ or V_h from which they are chosen is not subject to discussion. Neglecting bound constraints, the

parameter identification then has the form:

$$\min_{u\in\tilde{V},\ a_h\in\mathcal{A}_h} J(u,a_h), \qquad \text{subject to} \qquad (a_h\nabla u,\nabla\varphi)=(f,\varphi) \qquad \forall\varphi\in\tilde{V}.$$

In order to view discretization of a as a constraint, we first rewrite this minimization problem as one over a continuous space $\mathcal{A}$, but then again restrict a_h explicitly to $\mathcal{A}_h$. The above problem is then equivalent to finding $u\in\tilde{V}, a_h\in\mathcal{A}$ and numbers α_i such that

$$\begin{aligned} &\min_{u\in\tilde{V},\ a_h\in\mathcal{A}} J(u,a_h), \\ \text{subject to} \qquad &(a_h\nabla u,\nabla\varphi)=(f,\varphi) && \forall\varphi\in\tilde{V}, \\ &\langle a_h-\Sigma_i\alpha_i\chi_i,\eta\rangle=0 && \forall\eta\in\mathcal{A}', \end{aligned} \tag{2.12}$$

where the χ_i are the shape functions of $\mathcal{A}_h$. Introducing a Lagrange multiplier $\gamma\in\mathcal{A}'$ for the last constraint, the optimality system for this problem contains the equations of Problem 1.8, (1.3), but also

$$\langle\gamma,\chi_h\rangle=0 \qquad \forall\chi_h\in\mathcal{A}_h, \tag{2.13}$$

$$\nabla_a L(x;\chi)+\langle\gamma,\chi\rangle=0 \qquad \forall\chi\in\mathcal{A}, \tag{2.14}$$

with L the Lagrange functional already introduced in Problem 1.8. The Lagrange multiplier of the discretization constraint can thus be identified with the residual $-\nabla_a L$, which is orthogonal to $\mathcal{A}_h$ with respect to the duality pairing $\langle\cdot,\cdot\rangle$.

Note that by the reformulation, we have extended the test space for the equation concerning $\nabla_a L$ from $\mathcal{A}_h$ to $\mathcal{A}$ in (2.14). However, this increase is countered by the additional term in (2.14), which deletes that part of $\nabla_a L$ that is not orthogonal to the surplus test space $\mathcal{A}\backslash\mathcal{A}_h$.

Remark 2.5. *If we have discretized u,λ to u_h and λ_h, and if $r(a)=\frac{1}{2}\|a\|^2$, then by* (2.14) *we have the explicit representation*

$$\gamma=\beta a_h+\nabla u_h\cdot\nabla\lambda_h.$$

Since Lagrange multipliers represent the first order response of the objective function to a small change in the constraints, γ gives an indication how a relaxation of the discretization constraint would change the value of $J(u,a)$. Thus, if we would weaken the discreteness constraint in (2.12) to $a-\Sigma_i\alpha_i\chi_i=g$, then

$$J(x)-J(\tilde{x})=\langle g,\gamma\rangle+\mathcal{O}(\|g\|^2), \tag{2.15}$$

with $x,\tilde{x}$ the solutions of the original problem with discreteness constraint, and of the problem with perturbed constraint, respectively.

The mesh should then be refined in such a way that the objective function decreases maximally, which we assume to coincide with the best strategy for the identification of the unknown coefficient. Refining a cell then corresponds

to enriching the space $\mathcal{A}_h$ by the shape functions of $\mathcal{A}_{h/2}$, so we have to check the change in $J(\cdot)$ for functions $g \in \mathcal{A}_{h/2}$. As refinement indicator we then take

$$\eta^{\gamma}_{K_a} = \sum_i \left| \langle g^i_{K_a}, \gamma \rangle \right|, \tag{2.16}$$

where the $g^i_{K_a}$ form a basis of $\mathcal{A}_{h/2}$ on the cell K_a.

Remark 2.6. *If $\mathcal{A}_h = Q^0_d(\mathbb{T}_a)$, then we can choose the characteristic functions of the child cells K^c_a of K_a as basis of $\mathcal{A}_{h/2}$. Then,* (2.16) *reduces to*

$$\eta^{\gamma}_{K_a} = \sum_{K^c_a} \left| \int_{K^c_a} \gamma \, dx \right|.$$

Remark 2.7. *The refinement indicator $\eta^{\gamma}_{K_a}$ can be related to the residual $\rho^a_{K_a}$ from the approximate error estimate* (2.9)*. For example, for $\mathcal{A}_h = Q^0_d(\mathbb{T}_a)$ we have*

$$\eta^{\gamma}_{K_a} \leq \sqrt{|K_a|} \|\gamma\|_{K_a} = \sqrt{|K_a|} \rho^a_{K_a}.$$

The scaling factor equals the weight $\omega^a_{K_a} = h_{K_a} \|\nabla_h a_h\|$ if the coefficient is discontinuous since then $\omega^a_{K_a} \leq C\|a_h\| \leq C\sqrt{a_1}\sqrt{|K_a|}$.

Further exploiting the approach discussed above, we can derive a lower error bound for the coefficient from (2.15) under certain additional assumptions:

Theorem 2.8. *Let $x = \{u, a, \lambda\}$ and $x_h = \{u_h, a_h, \lambda_h\}$ be exact and discrete solutions. Assume that the state discretization allows to resolve state and dual variable exactly, and that for the error in the coefficient $\|a - a_h\|_{\mathcal{A}} < \delta$ with some fixed $\delta \geq 0$. Furthermore, assume that we have a lower estimate for the error in the objective functional, $\underline{\eta} \leq |J(x) - J(x_h)|$, then there exists a constant $C > 0$ such that*

$$\|a - a_h\|_{\mathcal{A}} \geq \frac{\underline{\eta} - C\delta^2}{\|\gamma\|_{\mathcal{A}'}}.$$

Proof. If we perturbed the discreteness constraint in (2.12) to $a - \Sigma_i \alpha_i \chi_i = g$ with $g = e_a = a - a_h$, then the exact solution a is on this constraint surface. The solution $\tilde{x}$ of the perturbed problem is thus the exact solution x since we have assumed that state and adjoint variable can be identified exactly. We then have by (2.15) that

$$\underline{\eta} \leq |J(x) - J(x_h)| \leq |\langle e_a, \gamma \rangle| + C\delta^2,$$

with some $C > 0$ bounding the higher order sensitivities in (2.15). The claim then follows by simple transformations. □

The relevance of the bound lies in the fact that as $\|a - a_h\| \to 0$, the quadratic term $C\delta^2$ on the right hand side tends to zero with higher order. This unknown second order term thus vanishes asymptotically.

By now, we have neglected the existence of bound constraints on a in the derivation of γ and $\eta^{\gamma}_{K_a}$. By (2.14), we know that γ can be expressed in terms of the residual $\nabla_a L(x)$, which should be zero for the exact continuous solution $x = \{u, a, \lambda\}$. However, if a is at one of its bounds, the gradient of the unconstrained Lagrangian is nonzero, but is countered by the Lagrange multiplier corresponding to this constraint, see (1.11). Thus, we change the definition of γ for the constrained problem to

$$\langle \tilde{\gamma}, \chi \rangle = \nabla_a L(x_h; \chi) \qquad \forall \chi \in \mathcal{A}(\Omega'), \tag{2.17}$$

where Ω' is the union of cells where the parameter is not at one of its bounds. For cells where the parameter is at either bound, $\tilde{\gamma}$ is extended by zero.

2.2.2 Criteria based on available information

Alternatively to (2.16), we have used other refinement criteria for the parameter mesh $\mathbb{T}_a$:

- If the dual weighted estimators (2.7) or (2.9) are used, we can use one of the following terms defined on the cells K_a of the parameter mesh for refinement:

$$\begin{aligned} \eta^{DWR1}_{K_a} &= \beta r'(a_h; \tilde{a} - i_h a) + (\nabla \lambda_h \cdot \nabla u_h, \tilde{a} - i_h a)_{K_a}, \\ \eta^{DWR2}_{K_a} &= \rho^a_{K_a} \tilde{\omega}^a_{K_a}. \end{aligned} \tag{2.18}$$

  Due to their derivation, we do not expect significant differences in their abilities as mesh refinement criteria and therefore only investigate the first one.

- If the state mesh was refined with one of the heuristic criteria defined in Section 2, then we can also use a more heuristic criterion for the refinement of the parameter mesh. For a piecewise constant approximation of the parameter we used

$$\eta^{\nabla a}_{K_a} = h^{1+d/2} \|\nabla_h a_h\|_{\infty;K}, \tag{2.19}$$

  where ∇_h is a difference quotient approximation to the gradient.

2.2.3 Comparison of refinement criteria

To assess the quality of the three refinement criteria $\eta^{\gamma}_{K_a}$ (2.16), $\eta^{DWR1}_{K_a}$ (2.18), and $\eta^{\nabla a}_{K_a}$ (2.19), we first look at the size of the refinement indicators for test case 2 (see page 37). Obviously, refinement should be directed entirely into the circular jump of the coefficient. Fig 2.7 shows the coefficient after the first few iterations on the initial mesh, as well as the distribution of the three indicators listed above.

In this case where the coefficient is well identified, all indicators roughly indicate the same cells for refinement. However, a common observation is that the DWR indicator only marks a very small number of cells for refinement,

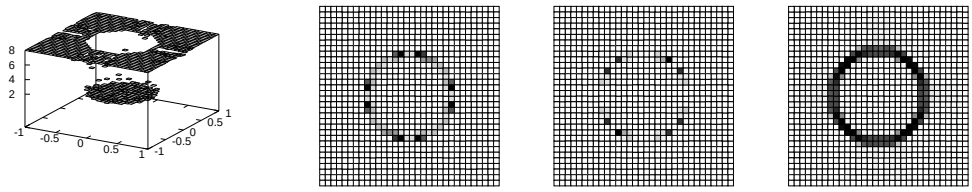


Figure 2.7: *Comparison of refinement criteria for the parameter mesh (test case 2). Left: Recovered parameter on coarse mesh. Center left: Values of* $\eta_{K_a}^{\gamma}$. *Center right: Values of* $\eta_{K_a}^{DWR1}$. *Right: Values of* $\eta_{K_a}^{\nabla a}$.

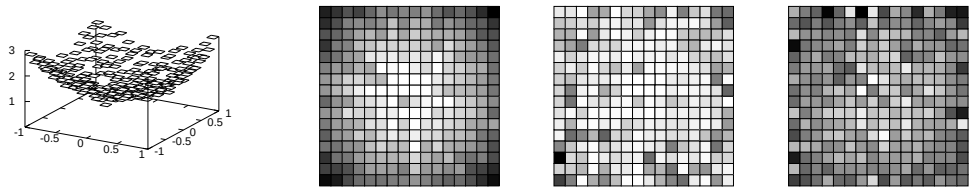


Figure 2.8: *Comparison of refinement criteria for the parameter mesh (test case 1 with added noise). Left: Recovered parameter on coarse mesh. Center left: Values of* $\eta_{K_a}^{\gamma}$. *Center right: Values of* $\eta_{K_a}^{DWR1}$. *Right: Values of* $\eta_{K_a}^{\nabla a}$.

leading to slow refinement of the parameter mesh. In addition, some refined cells are coarsened again in the next step. Thus, the DWR estimator often leads to rather unpredictable behavior unless a significant amount of heuristics are added. Due to this, no suitable refinement strategy could be found for some examples.

In contrast to this, $\eta_{K_a}^{\gamma}$ marks the cells around the circle in a more predictable way, while $\eta_{K_a}^{\nabla a}$ of course profits from the good approximation of the parameter and therefore has no problems indicating the correct cells.

As a second example, we consider the solution of test case 1 (see page 37), with 1.5% noise added. The presence of noise leads to a bad reconstruction of the parameter, which is amplified by the fact that we use piecewise constant elements for the parameter and only penalize the size, but not the roughness of the parameter by regularization.

The results of this experiment are shown in Fig. 2.8. While $\eta_{K_a}^{\gamma}$ seems relatively unaffected by the bad reconstruction and indicates those cells for refinement where the gradient of the *exact* solution is large (i.e. outwards from the center towards the corners) as should be expected, both the DWR and the ∇a indicator seems badly out of touch with the situation, proposing rather random cells for refinement.

As a summary, in general $\eta_{K_a}^{\gamma}$ is the most robust one, while $\eta_{K_a}^{DWR1}$ and $\eta_{K_a}^{\nabla a}$ were too unpredictable in their behavior and often suffered in the presence of noise. From the indicated relation between $\eta_{K_a}^{\gamma}$ and $\eta_{K_a}^{DWR1}$, it seems probable that the lacking robustness of the latter indicator is due to unreliable weights

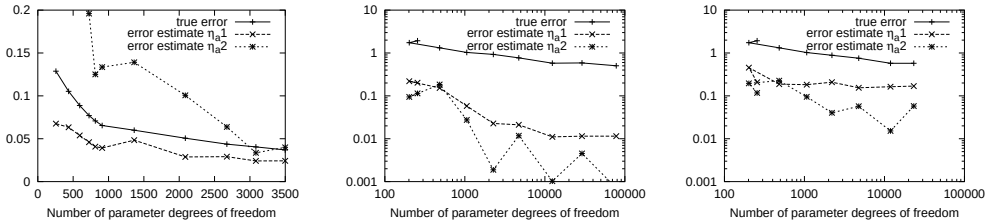


Figure 2.9: *Comparison of true error* $\|a_h - a_{exact}\|_{L^2}$ *and error estimates* η_a^1 *and* η_a^2. *Left: Test case 1. Center: Test case 2. Right: Test case 2, but with the estimate* $\tilde{\gamma}$ *incorporating bound constraints instead of the original* γ.

$\tilde{a} - a_h$, which is readily explained by lacking smoothness in a_h. If the application permits it, penalizing roughness might help in this case.

2.2.4 Reliability of error estimates

In order to check the accuracy of the lower error bound provided by Theorem 2.8, we consider the solution of two of the examples defined in Section 1.9. As a first test, we solve test case 1, with the regularization parameter $\beta = 0$ and no added noise. With these parameters, we know the exact solution a_{exact} of the problem and can compare the true error and the estimate. We also know the exact error in the functional J, since $J(x) = 0$ and thus $|J(x) - J(x_h)| = |J(x_h)|$. However, we do not know the value of the constant C appearing in the theorem. Neglecting this higher order term, we are led to compare $\|a_h - a_{exact}\|_{L^2}$ with the estimates

$$\eta_a^1 = |J(x_h)|/\|\gamma\|_{L^2}, \qquad \eta_a^2 = |\eta^{DWR1}|/\|\gamma\|_{L^2}.$$

The latter is computable even if the exact value of $J(x_h)$ is unknown. The true error in the coefficient $\|a_h - a_{exact}\|_{L^2}$, as well as the two estimates are reported in Fig. 2.9. It is seen that η_a^1 provides a reliable lower bound for the error. η_a^2 is too large at the beginning since η^{DWR1} initially overestimates the true error $J(x) - J(x_h)$.

In a second example, we take test case 2 to check the accuracy of the lower error bound. In contrast to the first example, here the bounds on the parameter are active in large areas of the domain. We thus expect that neglecting this fact in the derivation of γ will lead to an overestimated value of $\|\gamma\|$ and thus to an underestimated value of $\|a - a_h\|$. This can indeed be seen in the middle and right panel of Fig. 2.9, where the true error $\|a - a_h\|$ along with the estimates are shown that are obtained using γ and $\tilde{\gamma}$. It is clear that neglecting bound constraints leads to inefficient error bounds, while the estimate $\tilde{\gamma}$ incorporating bounds performs better.

2.3 Estimates based on stability

Besides the duality based strategies to get a posteriori error estimates, we briefly discuss another possibility for their construction. It is based on stability properties. The result contains a stability constant revealing the *worst case* stability of solutions instead of a dual solution representing the stability properties of a particular solution; the estimate will therefore greatly exceed the error in most cases. The construction of such an estimate is nevertheless shown as an alternative way.

Theorem 2.9. *Assume that the discretization space $\mathcal{X}_h = \mathcal{U}_h \times \mathcal{A}_h \times \Lambda_h$ admits the following interpolation estimate*

$$\inf_{y_h \in \mathcal{X}_h} \|y - y_h\| \le Ch\|y\|_{\mathcal{X}},$$

for all $y \in \mathcal{X}$, with $\|x\|_{\mathcal{X}}^2 = \|\nabla u\|_{L^2}^2 + \|a\|_{H^1}^2 + \|\nabla\lambda\|_{L^2}^2$. Assume further that the inf-sup condition

$$\sup_{y \in \mathcal{X}_0} \frac{A(x,y)}{\|y\|_{\mathcal{X}}} \ge \gamma \|x\|_{\mathcal{X}} \qquad \forall x \in \mathcal{X} \tag{2.20}$$

holds (see, e.g., Theorems 1.9 and 1.10), with $m(\varphi) = \frac{1}{2}\|\nabla\varphi\|^2$ and $r(\chi) = \frac{1}{2}\|\chi\|_{H^1}^2$, and

$$\begin{aligned} A(x,y) = (\nabla u, \nabla\varphi) + (a\nabla\lambda, \nabla\varphi) + (a\nabla u, \nabla\psi) \\ + \beta(a,\chi) + \beta(\nabla a, \nabla\chi) + (\nabla u \cdot \nabla\lambda, \chi). \end{aligned}$$

Let x^*, x_h^* be continuous and discrete solutions, respectively. Then the a posteriori estimate

$$\|e\|_{\mathcal{X}} \le \frac{C}{\gamma} \left\{ \sum_K h \left(\rho_K^u + \rho_K^a + \rho_K^\lambda \right) + h^{1/2} \left(\rho_{\partial K}^u + \rho_{\partial K}^a + \rho_{\partial K}^\lambda \right) \right\} + \mathcal{O}(\|e\|^2),$$

for the error $e = x^* - x_h^*$ holds with

$$\begin{aligned} \rho_K^u &= \|f + \nabla\cdot(a_h \nabla u_h)\|_K, & \rho_{\partial K}^u &= \|[\partial_n u_h]\|_{\partial K} + \|[a_h \partial_n u_h]\|_{\partial K} \\ \rho_K^\lambda &= \|\Delta(u_h - z) + \nabla\cdot(a_h\nabla\lambda_h)\|_K, & \rho_{\partial K}^\lambda &= \|[a_h \partial_n \lambda_h]\|_{\partial K}, \\ \rho_K^a &= \|\beta(a_h - \Delta a_h) + \nabla u_h \cdot \nabla\lambda_h\|_K, & \rho_{\partial K}^a &= \|[\partial_n a_h]\|_{\partial K}. \end{aligned}$$

Proof. Set $x = e = x^* - x_h^*$. Using Galerkin orthogonality, we have

$$\|e\|_{\mathcal{X}} \le \frac{1}{\gamma} \sup_{y \in \mathcal{X}} \frac{A(e,y)}{\|y\|_{\mathcal{X}}} = \frac{1}{\gamma} \sup_{y \in \mathcal{X}} \inf_{y_h \in \mathcal{X}_h} \frac{A(e, y - y_h)}{\|y\|_{\mathcal{X}}}.$$

Integrating by parts in $A(\cdot,\cdot)$, using the Cauchy-Schwarz inequality and the assumed interpolation estimate yields the following estimate (we drop the asterisk

on the elements of x^* and x_h^* for brevity):

$$\begin{aligned}\|e\|_{\mathcal{X}} \leq \frac{Ch}{\gamma}\Bigg\{ &\sum_K \| -\Delta u + \Delta u_h - \nabla\cdot(a\nabla\lambda) + \nabla\cdot(a_h\nabla\lambda) + \nabla\cdot(a\nabla\lambda_h) \\ &\qquad - \nabla\cdot(a_h\nabla\lambda_h)\|_K \\ &+ \sum_K \| - \nabla\cdot(a\nabla u) + \nabla\cdot(a_h\nabla u) + \nabla\cdot(a\nabla u_h) - \nabla\cdot(a_h\nabla u_h)\|_K \\ &+ \sum_K \|\beta(a - a_h - \Delta a + \Delta a_h) \\ &\qquad + \nabla u\cdot\nabla\lambda - \nabla u_h\cdot\nabla\lambda - \nabla u\cdot\nabla\lambda_h + \nabla u_h\cdot\nabla\lambda_h\|_K \\ &+ \sum_K \frac{1}{2}h^{-1/2}\Big(\|[\partial_n u_h]\|_{\partial K} + \|[a_h\partial_n u_h] - ([a_h\partial_n u] + [a\partial_n u_h])\|_{\partial K} \\ &\qquad + \|[a_h\partial_n\lambda_h] - ([a_h\partial_n\lambda] + [a\partial_n\lambda_h])\|_{\partial K} + \|[\partial_n a_h]\|_{\partial K}\Big)\Bigg\}.\end{aligned}$$

Using the optimality conditions for the continuous solution, we then obtain

$$\begin{aligned}\|e\|_{\mathcal{X}} \leq \frac{Ch}{\gamma}\Bigg\{ &\sum_K \|(-\Delta(u_h - z) - \nabla\cdot(a_h\nabla\lambda_h)) - (\nabla\cdot(e_a\nabla\lambda) + \nabla\cdot(a\nabla e_\lambda))\|_K \\ &+ \sum_K \|(-f - \nabla\cdot(a_h\nabla u_h)) - (\nabla\cdot(e_a\nabla u) + \nabla\cdot(a\nabla e_u))\|_K \\ &+ \sum_K \|(\beta(a_h - \Delta a_h) + \nabla u_h\cdot\nabla\lambda_h) + (\nabla e_u\cdot\nabla\lambda + \nabla u\cdot\nabla e_\lambda)\|_K \\ &+ \text{jump terms as above}\Bigg\}.\end{aligned}$$

By tangentiality, $\nabla\cdot(e_a\nabla\lambda) = -\nabla\cdot(a\nabla e_\lambda) + \mathcal{O}(\|e\|^2)$ and likewise for corresponding terms in the second parentheses of the other cell terms, as well as for the parentheses in the jump terms. We can thus split off these higher order terms and obtain the claimed result. □

Remark 2.10. *The result of Theorem 2.9 implicitly contains an estimate of the error in the parameter, since*

$$\|a - a_h\|_{H^1} \leq \|e\|_{\mathcal{X}}.$$

Nevertheless, the theorem is of little practical value since it incorporates the constant γ, denoting the worst case stability properties of $A(\cdot,\cdot)$. It does not, in general, reflect the stability of a particular solution and will thus lead to a large overestimation of the error. Exploiting the actual stability of a solution is only possible by taking into account the solution of a corresponding dual problem.

To illustrate the overestimation, note that in applications with small noise it is often possible to identify the parameter well with very small values of β; if, for example, $\beta = 10^{-8}$, then $\frac{1}{\gamma} > 10^8$.

2.4 Estimates for arbitrary functionals

In some cases, we may be interested in bounding the error with respect to functionals of the solution—including error functionals of the recovered parameter. In this section, we will derive an error representation for arbitrary functionals. First, we state the abstract form, only involving the Lagrangian of the problem, in Theorem 2.11, then apply it to the elliptic problem introduced in Chapter 1. Since the necessary computation of a dual quantity is too expensive for practical purposes, a modification is discussed that makes this feasible.

2.4.1 Statement of estimates

Theorem 2.11. *Let $E : \mathcal{X}_g \to \mathbb{R}$ be an error functional. Let $x \in \mathcal{X}_g$ be the solution of the stationarity condition $\nabla L(x; y) = 0$ for all $y \in \mathcal{X}_0$, and $\hat{x} \in \mathcal{X}_0$ be the solution of the dual problem*

$$\nabla_x^2 L(x; \hat{x}, y) = -\nabla_x E(x; y) \qquad \forall y \in \mathcal{X}_0. \tag{2.21}$$

Then the a posteriori error estimate

$$E(x) - E(x_h) = \frac{1}{2}\left\{\rho(x_h, \hat{x} - i_h\hat{x}) + \hat{\rho}(x_h, \hat{x}, x - i_h x)\right\} + R(x, x_h, \hat{x}, \hat{x}_h), \tag{2.22}$$

holds with residuals

$$\begin{aligned}\rho(x_h, y) &= \nabla_x L(x_h; y),\\ \hat{\rho}(x_h, \hat{x}, y) &= \nabla_x E(x_h; y) + \nabla_x^2 L(x_h; \hat{x}, y)\end{aligned}$$

and remainder term

$$\begin{aligned}R(x, x_h, \hat{x}, \hat{x}_h) = \tfrac{1}{2}\int_0^1 \Big\{&\nabla_x^3 E(x_h + se; e, e, e)\\ &+ \nabla_x^3 L(x_h + se; e, e, e) + \nabla_x^4 L(x_h + se; \hat{x} + s\hat{e}, e, e, e)\Big\} s(s-1)\, ds,\end{aligned}$$

where $e = x - x_h$, $\hat{e} = \hat{x} - \hat{x}_h$, and $\hat{x}_h$ the solution of a discrete counterpart of (2.21).

Proof. Let $\Xi = \{x, \hat{x}\} \in \mathcal{X}_g \times \mathcal{X}_0$, then x and $\hat{x}$ satisfy the identity

$$\nabla_\Xi \Lambda(\Xi; \xi) = 0 \qquad \forall \xi \in \mathcal{X}_0 \times \mathcal{X}_0,$$

with the joint Lagrangian $\Lambda(\Xi) = E(x) + \nabla_x L(x; \hat{x})$ containing the Lagrangian $L(x)$ of the first order conditions, see Problem 1.8. The proof continues in the same manner as the proof of Theorem 2.1, yielding

$$E(x) - E(x_h) = \tfrac{1}{2}\nabla_\Xi \Lambda(\Xi_h, e_\Xi) + \tfrac{1}{2}\int_0^1 \nabla_\Xi^3 \Lambda(\Xi_h + se_\Xi; e_\Xi, e_\Xi, e_\Xi)\; s(s-1)\, ds,$$

where $e_\Xi = \Xi - \Xi_h$ with $\Xi_h = \{x_h, \hat{x}_h\}$. The claim then follows by observing that

$$\nabla_\Xi \Lambda(\Xi_h; e_\Xi) = \nabla_x E(x_h, e) + \nabla_x^2 L(x_h; \hat{x}, e) + \nabla_x L(x_h; \hat{e}),$$

with $\hat{e} = \hat{x} - \hat{x}_h$, and using Galerkin orthogonality on these terms to replace e by $x - i_h x$, and $\hat{e}$ by $\hat{x} - i_h \hat{x}$. For the remainder term,

$$\begin{aligned}\nabla_\Xi^3 \Lambda(\Xi_h + s e_\Xi, e_\Xi, e_\Xi, e_\Xi) &= \nabla_x^3 E(x_h + se; e, e, e) \\ &\quad + \nabla_x^3 L(x_h + se; e, e, e) + \nabla_x^4 L(x_h + se; \hat{x} + s\hat{e}, e, e, e).\end{aligned}$$

□

For the particular elliptic problem considered here, the error estimate above, neglecting the remainder term R, assumes the following form:

$$\begin{aligned}E(x) - E(x_h) \approx \frac{1}{2}\Big\{ &\rho_u(x_h; \hat{x} - i_h\hat{x}) + \rho_a(x_h; \hat{x} - i_h\hat{x}) + \rho_\lambda(x_h; \hat{x} - i_h\hat{x}) \\ &+ \hat{\rho}_u(x_h; \hat{x}, \hat{x} - i_h\hat{x}) + \hat{\rho}_a(x_h; \hat{x}, \hat{x} - i_h\hat{x}) + \hat{\rho}_\lambda(x_h; \hat{x}, \hat{x} - i_h\hat{x})\Big\},\end{aligned}$$

with the residuals

$$\begin{aligned}\rho_u(x_h; \hat{x} - i_h\hat{x}) &= m'(u_h - z; \hat{u} - i_h\hat{u}) + (a_h \nabla\lambda_h, \nabla(\hat{u} - i_h\hat{u})) \\ \rho_a(x_h; \hat{x} - i_h\hat{x}) &= \beta r'(a_h; \hat{a} - i_h\hat{a}) + (\nabla\lambda_h \cdot \nabla u_h, \hat{a} - i_h\hat{a})), \\ \rho_\lambda(x_h; \hat{x} - i_h\hat{x}) &= (a_h \nabla u_h, \nabla(\hat{\lambda} - i_h\hat{\lambda})) - (f, \hat{\lambda} - i_h\hat{\lambda}), \\ \hat{\rho}_u(x_h; \hat{x}_h, x - i_h x) &= m''(u_h; \hat{u}_h, u - i_h u) + (\nabla\lambda_h \cdot \nabla\hat{u}_h, a - i_h a) \\ &\quad + (a_h \nabla\hat{u}_h, \nabla(\lambda - i_h\lambda)) + \nabla_u E(x_h; u - i_h u) \\ \hat{\rho}_a(x_h; \hat{x}_h, x - i_h x) &= \beta r''(a_h; \hat{a}_h, a - i_h a) + (\hat{a}_h \nabla\lambda_h, \nabla(u - i_h u)) \\ &\quad + (\hat{a}_h \nabla u_h, \nabla(\lambda - i_h\lambda)) + \nabla_a E(x_h; a - i_h a) \\ \hat{\rho}_\lambda(x_h; \hat{x}_h, x - i_h x) &= (a_h \nabla\hat{\lambda}_h, \nabla(u - i_h u)) + (\nabla u_h \cdot \nabla\hat{\lambda}_h, a - i_h a) \\ &\quad + \nabla_\lambda E(x_h; \lambda - i_h\lambda).\end{aligned}$$

For localized refinement criteria, these residuals should be evaluated only after cell-wise integration by parts, resulting in cell and face terms. The neglected remainder term has the form

$$R = -\frac{1}{12}((a-a_h)\nabla(\lambda-\lambda_h), \nabla(u-u_h)) + \frac{1}{2}\int_0^1 \nabla_x^3 E(x+se; e, e, e)\ s(s-1)\ ds.$$

In order to evaluate the error representation practically, we need the exact dual solution $\hat{x}$, or an approximation to it. Since the discrete counterpart of (2.21) is equivalent to one Newton step, the effort for the computation of some $\hat{x}_h$ equals the computation of one search direction for the full Newton method. Regarding the actual evaluation of the error estimate, the same possibilities exist as in Section 2.1.

Since the possibility of solving for exact Newton updates was already discarded for the solution of the inverse problems, the solution of (2.21) is too

expensive for the evaluation of an error estimate. Rather, we would like to use an approximate solution that satisfies a Gauß-Newton-type equation. The following theorem derives an estimate based on this idea:

Theorem 2.12. *With the same notation as in Theorem 2.11, split the Hessian as follows:*

$$\nabla_x^2 L(x;\hat{x},y) = H_1(x;\hat{x},y) + H_2(x;\hat{x},y),$$

where H_2 contains all second-order terms involving λ, and H_1 all other terms. Let now $\hat{x} \in \mathcal{X}_0$ be the solution of the Gauß-Newton system

$$H_1(x;\hat{x},y) = -\nabla_x E(x;y) \qquad \forall y \in \mathcal{X}_0. \tag{2.23}$$

Then there holds the a posteriori error estimate

$$E(x) - E(x_h) = \frac{1}{2}\left\{\rho(x_h, \hat{x} - i_h\hat{x}) + \hat{\rho}(x_h, \hat{x}, x - i_h x)\right\} + R'(x, x_h, \hat{x}, \hat{x}_h), \tag{2.24}$$

with remainder term

$$R'(x, x_h, \hat{x}, \hat{x}_h) = R(x, x_h, \hat{x}, \hat{x}_h) + \tfrac{1}{2}H_2(x;\hat{x},e),$$

where $e = x - x_h$ and $\hat{e} = \hat{x} - \hat{x}_h$, and residuals and remainder R as in Theorem 2.11.

Proof. With the same Lagrangian $\Lambda(\Xi)$ as in the proof of Theorem 2.11, we again have by approximation of the integral by the trapezoidal rule that

$$\begin{aligned} E(x) - E(x_h) = {} & \tfrac{1}{2}\nabla_\Xi \Lambda(\Xi, e_\Xi) + \tfrac{1}{2}\nabla_\Xi \Lambda(\Xi_h, e_\Xi) \\ & + \tfrac{1}{2}\int_0^1 \nabla_\Xi^3 \Lambda(\Xi_h + s e_\Xi; e_\Xi, e_\Xi, e_\Xi)\; s(s-1)\, ds. \end{aligned}$$

However, since $\hat{x}$ is now the solution of a perturbed problem, we no more have that $\nabla_\Xi \Lambda(\Xi;\xi) = 0$ for all test functions $\xi \in \mathcal{X}_0 \times \mathcal{X}_0$ so that the first term vanishes. Rather, we only have that

$$\begin{aligned} \nabla_{\hat{x}} \Lambda(\Xi; y) &= \nabla_x L(x; y) = 0 && \forall y \in \mathcal{X}_0, \\ \nabla_x \Lambda(\Xi; y) &= \nabla_x E(x; y) + \nabla_x^2 L(x;\hat{x},y) = H_2(x;\hat{x},y) && \forall y \in \mathcal{X}_0, \end{aligned}$$

by the decomposition of $\nabla_x^2 L$ defined above. The remainder R' is thus the sum of the previous remainder R and the new residual term involving H_2. □

The effort to obtain an approximation of the dual solution $\hat{x}$ used in this error identity is now equivalent to solving one additional Gauß-Newton step. Note that the main part of the error representation is the same as in the previous Theorem 2.11, only the remainder term changes.

For the elliptic equation considered so far, the residuals are those defined after Theorem 2.11, while the remainder term now has the additional part

$$H_2(x;\hat{x},e) = (\nabla\lambda, \hat{a}\nabla(u - u_h)) + (\nabla\lambda\cdot\nabla\hat{u}, a - a_h).$$

In the noise free case, if the measurement z is actually attainable, i.e. at the solution $u = z$, we have that $\lambda = 0$ and the additional term in the remainder vanishes.

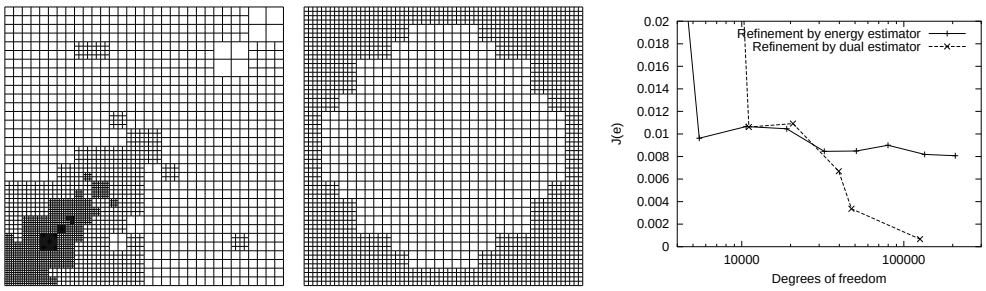


Figure 2.10: *Left: Mesh as produced by duality based estimate. Center: Mesh as produced by estimator with respect to $J(\cdot)$. Right: Error reduction for refinement by "energy indicator"* (2.7) *and by dual estimator* (2.24).

2.4.2 Results

In practical applications of distributed parameter estimation problems, the interesting quantities are usually values of the unknown coefficient at points or in subdomains. Since point values might not be defined properly, we replace them by mean values in a small neighborhood of the interesting point. We will therefore only consider examples of error functionals $E(\cdot)$ acting on $\{u, a, \lambda\}$ of the form

$$E(\{u, a, \lambda\}) = \int_\Omega \psi(\mathbf{x})\, a(\mathbf{x})\, dx,$$

where ψ is a weighting function.

Example 1. Consider test case 1 (see page 37) and assume we are interested in the value of the coefficient at the point $\mathbf{x}_0 = (-\frac{2}{3}, -\frac{2}{3})$. Using $\varepsilon = 0.05$, we set the weighting function to

$$\psi(\mathbf{x}) = \begin{cases} 1/(\pi\varepsilon^2) & \text{if } |\mathbf{x} - \mathbf{x}_0| < \varepsilon, \\ 0 & \text{otherwise.} \end{cases}$$

Fig. 2.10 shows typical meshes as produced by the duality error representation of Theorem 2.12 with respect to the functional $E(\cdot)$, and by the estimate (2.7) with respect to $J(\cdot)$. While the latter mostly sees the uniformly good approximation of a quadratic function by bilinear elements on a globally refined mesh, the former adapts the mesh towards the evaluation point $(-\frac{2}{3}, -\frac{2}{3})$. The figure also shows the superiority as refinement criterion of the dual estimator (2.24) over the "energy indicator" (2.7).

Example 2. In order to check the accuracy of (2.24) for the actual estimation of errors, we consider a more challenging example: take test case 4 (page 39) and as target functional use the mean error in the left sector, which is characterized by the weight

$$\psi(\mathbf{x}) = \begin{cases} 1 & \text{if } \frac{x_1}{3} \le x_2 < -\frac{x_1}{3}, x_1 < 0, \\ 0 & \text{otherwise.} \end{cases}$$

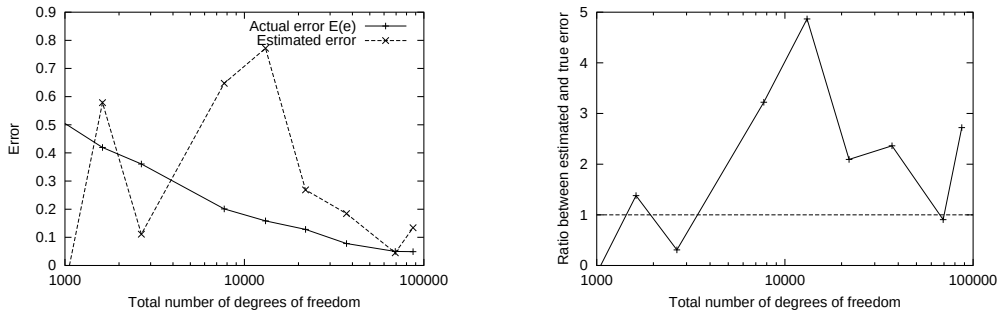


Figure 2.11: *Left: Comparison of actual error* $E(x - x_h)$ *and estimate* (2.24). *Right: Overestimation ratio.*

Fig. 2.11 shows that the error estimates have the correct order of magnitude, but are not quite accurate. This also holds for other examples. The reason for this is presently unclear.

2.5 Estimates for the constrained problem

In this chapter, we have up to now derived error estimates for Problem 1.7 under the assumption that inequality constraints can be neglected. We will extend these estimates to the constrained case in this section.

The basic problem in the incorporation of inequality constraints is that the first order necessary conditions as stated in Problem 1.14 include inequality constraints as well, although only implicitly in the definition of the dual cone $\mathcal{C}^+$ in (1.10) from which the respective Lagrange multipliers are chosen. The solution is thus characterized by a variational inequality.

A general framework for error estimation for variational inequalities has been proposed by Suttmeier and Blum, see [21, 62, 61]. Although we obtain related results, we will rather derive estimates for this problem by reformulating it as an equality constrained one which we obtain by presuming that we know the regions of the domain where the coefficient is at its bounds. For this, define by

$$\mathcal{I}^0 = \{\mathbf{x} : a(\mathbf{x}) = a_0\}, \qquad \mathcal{I}^1 = \{\mathbf{x} : a(\mathbf{x}) = a_1\},$$

the sets where the exact solution a is at its bounds. Likewise, let

$$\mathcal{I}_h^0 = \{\mathbf{x} : a_h(\mathbf{x}) = a_0\}, \qquad \mathcal{I}_h^1 = \{\mathbf{x} : a_h(\mathbf{x}) = a_1\},$$

be the sets where the numerical approximation is at its bounds. With this, define a Lagrangian by

$$\mathcal{L}(x, \mu_0, \mu_1; S^0, S^1) = L(x) + \langle \mu_0, a - a_0 \rangle_{S^0} + \langle \mu_1, a_1 - a \rangle_{S^1}, \tag{2.25}$$

where L is the original Lagrangian as defined in Problem 1.7, and S^i are sets where constraints $a = a_i$ will be prescribed. Then continuous and discrete

solutions trivially satisfy the stationarity conditions of problems with *equality* constraints on the parameter:

$$\begin{aligned} \nabla_x \mathcal{L}(x, \mu_0, \mu_1; \mathcal{I}^0, \mathcal{I}^1; y) = 0 \qquad & \forall y \in \mathcal{X}_0, \\ \langle \gamma_i, a - a_i \rangle_{\mathcal{I}^i} = 0 \qquad & \forall \gamma_i \in L^1, i = 0, 1, \end{aligned} \tag{2.26}$$

and

$$\begin{aligned} \nabla_x \mathcal{L}(x_h, \mu_{0,h}, \mu_{1,h}; \mathcal{I}_h^0, \mathcal{I}_h^1; y_h) = 0 \qquad & \forall y_h \in \mathcal{X}_h, \\ \langle \gamma_{i,h}, a_h - a_i \rangle_{\mathcal{I}_h^i} = 0 \qquad & \forall \gamma_{i,h} \in \mathcal{A}_h, i = 0, 1. \end{aligned} \tag{2.27}$$

Here, the Lagrange multipliers are discretized by the same spaces as the parameters, and the active sets implicitly depend on the solution. Since the Lagrange multipliers are defined only on the active sets, we are free to extend them by zero to the whole domain.

2.5.1 Estimates for the minimization functional

With the conditions above, we first derive the following a posteriori estimate with respect to the functional $J(\cdot)$ for the bound constrained problem. An intuitive interpretation is given afterwards.

Theorem 2.13. *Let $\xi = \{x, \mu^i\}$ and $\xi_h = \{x_h, \mu_h^i\}, i = 1, 2$, be the solutions of the inequality constrained problems* (2.26) *and* (2.27). *Define by*

$$\mathcal{I}_+^i = \mathcal{I}^i \backslash \mathcal{I}_h^i, \qquad \mathcal{I}_-^i = \mathcal{I}_h^i \backslash \mathcal{I}^i, \qquad i = 1, 2$$

that parts of the continuous and discrete active set that are not in the common subset of the two. Then there holds the error representation

$$\begin{aligned} J(x) - J(x_h) = \frac{1}{2} \left[\nabla_x L(x_h; x - y_h) + \langle \mu_{0,h}, a - \chi_h \rangle_{\mathcal{I}_h^0} - \langle \mu_{1,h}, a - \chi_h \rangle_{\mathcal{I}_h^1} \right] \\ + Q + R, \end{aligned} \tag{2.28}$$

for all $y_h = \{\varphi_h, \chi_h, \psi_h\} \in \mathcal{X}_h$, with

$$\begin{aligned} Q = \tfrac{1}{2} \left\{ \langle \mu_{0,h}, a - a_0 \rangle_{\mathcal{I}_-^0} - \langle \mu_0, a_h - a_0 \rangle_{\mathcal{I}_+^0} \right\} \\ - \tfrac{1}{2} \left\{ \langle \mu_{1,h}, a - a_1 \rangle_{\mathcal{I}_-^1} - \langle \mu_1, a_h - a_1 \rangle_{\mathcal{I}_+^1} \right\}, \end{aligned}$$

and the nonlinear remainder R as in Theorem 2.1:

$$R = -\frac{1}{12} ((a - a_h) \nabla(\lambda - \lambda_h), \nabla(u - u_h)).$$

Proof. As upper and lower bounds are treated in exactly the same way, we only show the proof of the theorem for the terms involving the lower constraint and denote the Lagrange multiplier for this constraint by $\mu = \mu_0$, the active set by $\mathcal{I} = \mathcal{I}^0$, and likewise for $\mathcal{I}_h, \mathcal{I}_+, \mathcal{I}_-$. The derivation of the respective terms for the upper constraint is straightforward based on the proof.

Since at the solutions ξ, ξ_h, state equation and bounds are satisfied with respect to corresponding test spaces, we have that

$$\begin{aligned} J(\xi) - J(\xi_h) &= \mathcal{L}(\xi;\mathcal{I}) - \mathcal{L}(\xi_h;\mathcal{I}_h) \\ &= \underbrace{\mathcal{L}(\xi;\mathcal{I}_h) - \mathcal{L}(\xi_h;\mathcal{I}_h)}_{A_1} + \underbrace{\mathcal{L}(\xi;\mathcal{I}) - \mathcal{L}(\xi;\mathcal{I}_h)}_{A_2}. \end{aligned}$$

The two parts are treated separately. For the first one, all integrals extend over the same domains. Denoting $e_\xi = \xi - \xi_h$, we have by the same argument used for the other estimates that

$$A_1 = \tfrac{1}{2}\nabla_\xi \mathcal{L}(\xi;\mathcal{I}_h;e_\xi) + \tfrac{1}{2}\nabla_\xi \mathcal{L}(\xi_h;\mathcal{I}_h;e_\xi) + \tfrac{1}{2}\int_0^1 \nabla_\xi^3 \mathcal{L}(\xi_h;\mathcal{I}_h;e_\xi,e_\xi,e_\xi)s(s-1)ds.$$

Since the bounds terms in $\mathcal{L}$ are only quadratic in the variables, the third derivative of $\mathcal{L}$ equals the third derivative of L, yielding the remainder term R. For the first term, we use the stationarity condition (2.26) to cancel the terms involving domain integrals and to separate the integrals over the active sets into different parts to obtain

$$\begin{aligned} \nabla_\xi \mathcal{L}(\xi;\mathcal{I}_h;e_\xi) &= \nabla_\xi \left[\mathcal{L}(\xi;\mathcal{I}) + \langle \mu, a - a_0\rangle_{\mathcal{I}} - \langle \mu, a - a_0\rangle_{\mathcal{I}_h}\right](e_\xi) \\ &= \nabla_\xi \left[\langle \mu, a - a_0\rangle_{\mathcal{I}_+} - \langle \mu, a - a_0\rangle_{\mathcal{I}_-}\right](e_\xi). \end{aligned}$$

Using that $a|_{\mathcal{I}_+} = a_0$, $a_h|_{\mathcal{I}_-} = a_0$, and $\mu|_{\mathcal{I}_-} = 0$, this term further reduces to

$$\nabla_\xi \mathcal{L}(\xi;\mathcal{I}_h;e_\xi) = -\langle \mu, a_h - a_0\rangle_{\mathcal{I}_+} + \langle \mu_h, a - a_0\rangle_{\mathcal{I}_-}$$

Likewise, we find

$$\nabla_\xi \mathcal{L}(\xi_h;\mathcal{I}_h;e_\xi) = \nabla_x L(x_h;e) + \langle \mu_h, a - a_h\rangle_{\mathcal{I}_h} + \langle \mu - \mu_h, a_h - a_0\rangle_{\mathcal{I}_h},$$

where the last term vanishes. By the first optimality condition in (2.27), we can replace the weight $\hat{e}$ by any $x - y_h$ for $y_h \in \mathcal{X}_h$.

The second term A_2, using cancellation, reduces to integrals over the active sets. Again noting that $a|_{\mathcal{I}} = a_0$, $\mu|_{\mathcal{I}_-} = 0$, we have

$$A_2 = \langle \mu, a - a_0\rangle_{\mathcal{I}} - \langle \mu, a - a_0\rangle_{\mathcal{I}_h} = 0.$$

Putting it all together, and treating the terms due to the upper bound alike, we obtain the claimed result. □

2.5.2 Interpretation and evaluation

The error representation derived above has an intuitive interpretation. First, note that if we identified the active set correctly, i.e. $\mathcal{I}^i = \mathcal{I}_h^i$, $i = 1, 2$, then the term denoted by Q vanishes since $\mathcal{I}_\pm^i \equiv \emptyset$. For this case,

$$\nabla_x \mathcal{L}(x_h, \mu_{0,h}, \mu_{1,h};\cdot) = \nabla_x L(x_h;\cdot) + \langle \mu_{0,h}, \cdot\rangle_{\mathcal{I}^0} - \langle \mu_{1,h}, \cdot\rangle_{\mathcal{I}^1}$$

is the residual of the first optimality condition in (2.26). As usual in a posteriori energy estimates, this residual is weighted by some $x - y_h$ with an arbitrary $y_h \in \mathcal{X}_h$.

In the other case, when we have not identified the active sets correctly, the term Q does not vanish. However, it is quadratic in the error: for example, for the first term in Q note that $\mu_0|_{\mathcal{I}_-^0} = 0$ and $a_h|_{\mathcal{I}_-^0} = a_0$. Thus

$$\langle \mu_{0,h}, a - a_0 \rangle_{\mathcal{I}_-^0} = - \langle \mu_0 - \mu_{0,h}, a - a_h \rangle_{\mathcal{I}_-^0} .$$

We may thus neglect it and only work with the main part of the error representation.

In Chapter 3, methods for the actual inclusion of bounds into the solution process are discussed. The method of choice there is an active set method which includes estimates for the Lagrange multipliers without explicitly computing them. The evaluation of the error representation above is therefore not straightforward since the μ_h^i are lacking. The following lemma states that the evaluation is possible nevertheless:

Lemma 2.14. *Denote by $g_h \in \mathcal{A}_h$ the discrete projection of $\nabla_a L(x_h; \cdot)$, i.e.*

$$(g_h, \chi_h) = \nabla_a L(x_h; \chi_h) \qquad \forall \chi_h \in \mathcal{A}_h.$$

Then the main part of the error representation in Theorem 2.13 can be written as

$$\begin{aligned} &\nabla_x L(x_h; x - y_h) + \langle \mu_{0,h}, a - \chi_h \rangle_{\mathcal{I}_h^0} - \langle \mu_{1,h}, a - \chi_h \rangle_{\mathcal{I}_h^1} \\ &\quad = \nabla_u L(x; u - \varphi_h) + \nabla_a L(x; a - \chi_h) - (g_h, a - \chi_h) + \nabla_\lambda L(x; \lambda - \psi_h). \end{aligned}$$

Proof. Since the Lagrange multipliers μ_h^i are only defined on the discrete active sets, we can define

$$\mu_h(\mathbf{x}) = \begin{cases} \mu_{0,h} & \text{for } \mathbf{x} \in \mathcal{I}_h^0, \\ -\mu_{1,h} & \text{for } \mathbf{x} \in \mathcal{I}_h^1, \\ 0 & \text{elsewhere.} \end{cases}$$

Selecting now the a-derivative in the optimality condition (2.27), we have that

$$\begin{aligned} 0 = \nabla_a \mathcal{L}(x_h, \mu_{0,h}, \mu_{1,h}; \mathcal{I}_h^0, \mathcal{I}_h^1; y_h) &= \nabla_a L(x; \chi_h) + \langle \mu_{0,h}, \chi_h \rangle_{\mathcal{I}_h^0} - \langle \mu_{1,h}, \chi_h \rangle_{\mathcal{I}_h^1} \\ &= \nabla_a L(x; \chi_h) + \langle \mu_h, \chi_h \rangle \end{aligned}$$

for all discrete test functions $\chi_h \in \mathcal{A}_h$. We thus see that $g_h = -\mu_h$ and

$$\langle \mu_{0,h}, a - \chi_h \rangle_{\mathcal{I}_h^0} - \langle \mu_{1,h}, a - \chi_h \rangle_{\mathcal{I}_h^1} = \langle \mu_h, a - \chi_h \rangle = - (g_h, a - \chi_h) .$$

□

The result shows that even if we did not compute the Lagrange multipliers, the error estimate can be evaluated: the missing multipliers can be obtained by projection of the gradient of the Lagrangian. Since this projection is local for the discontinuous shape functions of $\mathcal{A}_h$, and since $g_h = -\mu_h$ is zero outside the active sets, this is cheap.

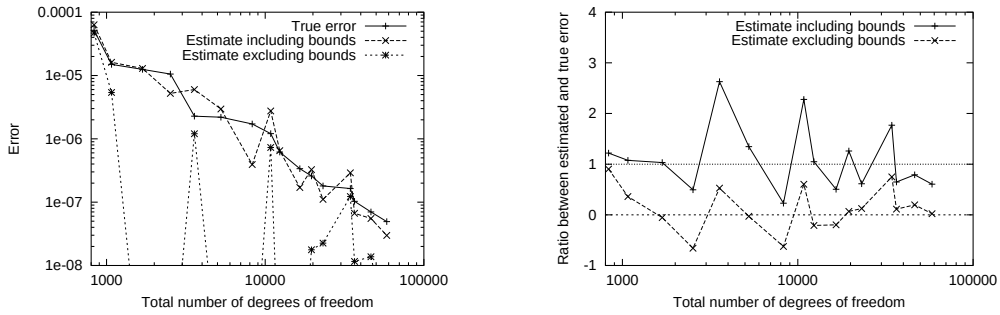


Figure 2.12: *Comparison of true error, and error estimates including bound constraints* (2.29)*, and neglecting these constraints* (2.7)*. Left: Values of error and estimates; values with the wrong sign are plotted at* $-\infty$*. Right: Ratio between estimated and true error.*

Remark 2.15. *If the linear functional* $\nabla_a L(x;\cdot)$ *allows a Riesz representation* $g \in L^1 = \mathcal{A}'$*, then the identity in Lemma 2.14 can be written in a way more appealing to intuition as*

$$\begin{aligned}\nabla_x L(x_h; x - y_h) + \langle \mu_{0,h}, a - \chi_h \rangle_{\mathcal{I}_h^0} - \langle \mu_{1,h}, a - \chi_h \rangle_{\mathcal{I}_h^1} \\ = \nabla_u L(x; u - \varphi_h) + (g - P_h g, a - \chi_h) + \nabla_\lambda L(x; \lambda - \psi_h),\end{aligned}$$

where $P_h g$ *is the projection of* g *onto* $\mathcal{A}_h$*. In some cases, this representation* g *can be obtained simply. For example, if* $r(a) = \frac{1}{2}\|a\|^2$*, then the Riesz representation* g *of* $\nabla_a L(x_k;\cdot)$ *is*

$$g = \beta a_k + \nabla u_k \cdot \nabla \lambda_k.$$

Given the above considerations, localized refinement criteria can be obtained from the error representation using the same method as in Section 2.1, i.e. by cell wise integration by parts of the individual terms and approximation of the weights using the discrete solution.

2.5.3 Reliability of estimates

In this section, we assess the quality of error estimates based on Theorem 2.13 and Remark 2.15. We only consider the main part of the error representation in Theorem 2.13, i.e.

$$\eta = \frac{1}{2}\left\{\nabla_u L(x; u - \varphi_h) + (g - P_h g, a - \chi_h) + \nabla_\lambda L(x; \lambda - \psi_h)\right\}, \qquad (2.29)$$

where the individual terms are expanded into cell and face contributions as shown in Theorem 2.2.

Fig. 2.12 shows the value of this estimate compared to the true error for test case 2, where the exact coefficient is at either bound everywhere. Also shown is the value of the error estimate (2.7) which was derived under the assumption of no bound constraints.

While the estimate neglecting bounds does not accurately track the error and even mispredicts its sign, the estimate including bounds is relatively accurate, with overestimation factors bounded and in the range $\frac{1}{2} \dots 3$.

2.5.4 Estimates for arbitrary functionals

The ideas used in Sections 2.4 and 2.5.1 can be combined to obtain an error representation for the constrained problem with respect to arbitrary functionals:

Theorem 2.16. *Let $E : \mathcal{X}_g \to \mathbb{R}$ be an error functional acting on $x = \{u, a, \lambda\}$, i.e. it does not evaluate the Lagrange multipliers μ_i of the bound constraints. Let $\xi = \{x, \mu^i\}$ and $\xi_h = \{x_h, \mu_h^i\}, i = 1, 2$, be the solutions of the inequality constrained problems (2.26) and (2.27), and define by $\hat{\xi}, \hat{\xi}_h$ the solutions of the dual problems*

$$\nabla_\xi^2 \mathcal{L}(\xi; \mathcal{I}^0, \mathcal{I}^1; \hat{\xi}, \eta) = -\nabla_\xi E(\xi; \eta) \qquad \forall \eta \in \mathcal{X}_0 \times L^1 \times L^1, \tag{2.30}$$

$$\nabla_\xi^2 \mathcal{L}(\xi_h; \mathcal{I}^0, \mathcal{I}^1; \hat{\xi}_h, \eta_h) = -\nabla_\xi E(\xi_h; \eta_h) \qquad \forall \eta_h \in \mathcal{X}_h \times \mathcal{A}_h \times \mathcal{A}_h, \tag{2.31}$$

with $\mathcal{L}$ as defined in (2.25). Then the a posteriori error estimate

$$E(x) - E(x_h) = \frac{1}{2} \Big\{ \rho(\xi_h, \hat{\xi} - i_h \hat{\xi}) + \hat{\rho}(\xi_h, \hat{\xi}, \xi - i_h \xi) \Big\} + Q + R, \tag{2.32}$$

holds with residuals

$$\begin{aligned} \rho(\xi_h, \eta) &= \nabla_\xi \mathcal{L}(\xi_h; \eta), \\ \hat{\rho}(\xi_h, \hat{\xi}, \eta) &= \nabla_\xi E(\xi_h; \eta) + \nabla_\xi^2 L(\xi_h; \hat{\xi}, \eta), \end{aligned}$$

and remainder terms

$$\begin{aligned} Q(\xi, \xi_h, \hat{\xi}, \hat{\xi}_h) = {} & \tfrac{1}{2} \Big\{ \langle \hat{\mu}_0, a_h - a_0 \rangle_{\mathcal{I}_+^0} - \langle \hat{\mu}_{0,h}, a - a_0 \rangle_{\mathcal{I}_-^0} + \langle \mu_0, \hat{a}_h \rangle_{\mathcal{I}_+^0} - \langle \mu_{0,h}, \hat{a} \rangle_{\mathcal{I}_-^0} \Big\} \\ & - \tfrac{1}{2} \Big\{ \langle \hat{\mu}_1, a_h - a_1 \rangle_{\mathcal{I}_+^1} - \langle \hat{\mu}_{1,h}, a - a_1 \rangle_{\mathcal{I}_-^1} + \langle \mu_1, \hat{a}_h \rangle_{\mathcal{I}_+^1} - \langle \mu_{1,h}, \hat{a} \rangle_{\mathcal{I}_-^1} \Big\} \end{aligned}$$

and

$$\begin{aligned} R(x, x_h, \hat{x}, \hat{x}_h) = {} & \tfrac{1}{2} \int_0^1 \Big\{ \nabla_x^3 E(x_h + se; e, e, e) \\ & + \nabla_x^3 L(x_h + se; e, e, e) + \nabla_x^4 L(x_h + se; \hat{x} + s\hat{e}, e, e, e) \Big\} s(s-1)\, ds, \end{aligned}$$

where $e = x - x_h$, $\hat{e} = \hat{x} - \hat{x}_h$.

Proof. We use the same techniques as in the proofs of Theorems 2.11 and 2.13. Steps that were already performed there are not discussed again. For simplicity, we again restrict attention to the lower bounds and denote $\mu = \mu_0$, etc as in the proof of Theorem 2.13. The terms due to the upper bound can easily be added.

Let $\Xi = \{x, \mu, \hat{x}, \hat{\mu}\}$. Then continuous and discrete primal and dual solutions are solutions to

$$\nabla_\Xi \Lambda(\Xi; \mathcal{I}; \xi) = 0, \qquad \nabla_\Xi \Lambda(\Xi_h; \mathcal{I}_h; \xi_h) = 0,$$

for all continuous and discrete test functions ξ, ξ_h, with the joint Lagrangian

$$\Lambda(\Xi; S) = E(x) + \nabla_x \mathcal{L}(x; S; \hat{x}),$$

where the Lagrangian $\mathcal{L}$ as defined in (2.25). Then,

$$\begin{aligned} E(x) - E(x_h) &= \Lambda(\Xi; \mathcal{I}) - \Lambda(\Xi_h; \mathcal{I}_h) \\ &= \underbrace{\Lambda(\Xi; \mathcal{I}_h) - \Lambda(\Xi_h; \mathcal{I}_h)}_{A_1} + \underbrace{\Lambda(\Xi; \mathcal{I}) - \Lambda(\Xi; \mathcal{I}_h)}_{A_2}. \end{aligned}$$

The integrals in the term denoted by A_1 extend over the same domains and can be transformed as in all previous examples to yield

$$\begin{aligned} A_1 = \tfrac{1}{2} \underbrace{\nabla_\Xi \Lambda(\Xi; \mathcal{I}_h; e_\Xi)}_{B_1} + \tfrac{1}{2} \underbrace{\nabla_\Xi \Lambda(\Xi_h; \mathcal{I}_h; e_\Xi)}_{B_2} \\ + \underbrace{\tfrac{1}{2} \int_0^1 \nabla_\Xi^3 \Lambda(\Xi_h + s e_\Xi; \mathcal{I}_h; e_\Xi, e_\Xi, e_\Xi) s(s-1)\, ds}_{B_3}. \end{aligned}$$

The four terms A_2, B_1, B_2, B_3 will now be discussed separately.

First, expanding A_2 yields

$$A_2 = \langle \hat{\mu}, a - a_0 \rangle_{\mathcal{I}_+} - \langle \hat{\mu}, a - a_0 \rangle_{\mathcal{I}_-} + \langle \mu, \hat{a} \rangle_{\mathcal{I}_+} - \langle \mu, \hat{a} \rangle_{\mathcal{I}_-}$$

Since $a|_{\mathcal{I}_+} = a_0$ and μ can be extended by zero to $\mathcal{I}_-$, the first, second, and fourth term vanish. Using the defining equations for the dual solution $\hat{x}$, we see that $\hat{a}|_{\mathcal{I}_+} = 0$ if as assumed $E(x)$ does not depend on μ; the fourth term thus vanishes as well.

As in previous proofs, the terms B_1, B_2 and B_3 yield the term Q, the main part of the error representation, and the remainder term R, respectively. □

Regarding the evaluation of this error representation, or of its main part for practical purposes, the same possibilities exist as discussed in Sections 2.4 and 2.5.1. In particular, the use of a nearby Gauß-Newton system for the dual solution instead of the full Newton system is possible, resulting in the same additional term H_2 as in Theorem 2.12.

2.6 Practical aspects of mesh refinement

In this chapter, a number of a posteriori estimates have been derived. Besides some that used other techniques, we presented several that were derived using Galerkin orthogonality and the Lagrangian structure of the problem:

- (2.4) for the error with respect to the minimization functional $J(\cdot)$;
- (2.22) and (2.24) for the error with respect to arbitrary functionals $E(\cdot)$ of the solution;
- (2.28) for the error with respect to $J(\cdot)$ of the bound constrained problem;

- (2.32) for arbitrary functionals for the constrained problem.

These estimates had in common that their practical evaluation involves integrating by parts the given terms, thus splitting the estimates into cell and face residuals. These residuals are either weighted by a quantity derived from the solution itself (in case of estimates for $J(\cdot)$) or from the solution of a dual problem (in case of estimates for arbitrary functionals). The process of integrating by parts and splitting into different terms has been made explicit for the first estimate above in Section 2.1.1 and exemplary in Theorem 2.2. For all other estimates, this process is implied and necessary for useful estimates that can also be used for refinement.

After splitting the estimates into cell-wise terms, we obtain sums over the cells of the state mesh $\mathbb{T}$ and of the mesh used for the parameter discretization $\mathbb{T}_a$. These terms are not split up arbitrarily but rather possess a natural association with either of these meshes. It is readily seen that coarsening of one mesh does not imply that the quantities on the other mesh generate a larger residual, and vice versa for refinement. Therefore, the resulting refinement criteria for the two meshes are independent of each other.

Except for the cases discussed in Section 2.2.3 where the actual evaluation of the estimates including the approximation of weight factors presented some difficulties, the estimates listed above can therefore be used to drive refinement of both meshes without additional heuristics.

Chapter 3

Bound constraints on the parameters

In this chapter, we will discuss methods to enforce bound constraints

$$a_0 \leq a(\mathbf{x}) \leq a_1.$$

Of course, at least guaranteeing a lower bound $0 < \alpha \leq a(\mathbf{x})$ is essential for the well-posedness of the continuous problem, but enforcement of bounds with physical values a_0, a_1 is an important goal when trying to identify parameters that actually bear physical meaning.

Within this chapter, we will first discuss a successful method – a modified active set strategy – to enforce these bounds, then briefly mention two methods – transformation and projection – that did not work as well as expected. An application is shown at the end.

3.1 Treating parameter bounds by active sets

One very successful approach to treating inequality constraints in finite dimensional optimization is the use of so-called *active set methods*. In this section, we propose an active set strategy that differs from the usual methods (see, e.g., Nocedal and Wright [51]) in two respects:

- It scales the Lagrange multipliers in accordance with the size of the cells on which they are defined. This allows to view them as discretized versions of a continuous function, and avoids ill-conditioned problems for locally refined meshes.
- It modifies the strategy by which the active set is determined, guaranteeing the efficiency of the method.

Active set methods work by identifying a set of *active* constraints in each nonlinear step that are then considered as *equalities*. If this set is chosen appropriately, then it can be guaranteed that the set of constraints that are active at the solution is identified at some point in the process. Choosing the active set is not complicated and can be done using a simple preprocessing step before

each iteration and is particularly cheap if, as is the case here, the constraints are simple bounds.

In order to describe the active set strategy that was used, we briefly review how active sets work in the finite dimensional case first, then how they can be defined in the setting of a finite element discretization of continuous problems.

The finite dimensional case

Active set strategies for finite dimensional problems with inequality constraints $c_i(x) \geq 0$ are based on the observation that if $x^* \in \mathbb{R}^n$ is a local solution of the inequality constrained problem

$$\min_x f(x), \qquad \text{such that } c_i(x) \geq 0,\ i \in \mathcal{I},$$

then it trivially is also a solution of the following problem:

$$\min_x f(x), \qquad \text{such that } c_i(x) = 0,\ i \in \mathcal{I}_a \subset \mathcal{I}, \tag{3.1}$$

where the *active set* $\mathcal{I}_a$ of constraints is defined by $\mathcal{I}_a = \{i \in \mathcal{I} : c_i(x^*) = 0\}$. If we knew the active set $\mathcal{I}_a$, we could restate the inequality constrained problem as an equality constrained one. Unfortunately, the active set depends implicitly on the unknown exact solution x^*. Active set methods therefore work with approximations $W_k \subset \mathcal{I}$ to the exact active set $\mathcal{I}_a$, and try to make sure that $W_k \to \mathcal{I}_a$.

In order to derive an algorithm by which we can identify W_k for the special application discussed in this work, let $x = \{u, a, \lambda\}$ and consider the Lagrangian for the constrained discrete problem,

$$L_c(x, \mu) = L(x) + \mu^T c(a),$$

where $L(x)$ is the Lagrangian of the problem without inequalities, and $c_i(a) = a_i - a_0$ (for simplicity, we only consider lower bounds). Then the optimality condition includes the equations

$$\nabla_a L_c(x^*, \mu^*) = \nabla_a L(x^*) + \nabla c(a^*)^T \mu^* = 0, \qquad \mu_i \leq 0. \tag{3.2}$$

Due to the special structure of the bound constraints, $\nabla c = \mathbf{1}$, and for the optimal Lagrange multiplier

$$\mu^* = -\nabla_a L(x^*) \tag{3.3}$$

holds. If we have not yet found the optimum, the residual of the first equation in (3.2) will in general not vanish. However, we can define by

$$\mu_k = -\nabla_a L(x_k) \tag{3.4}$$

an approximation to the exact Lagrange multiplier. Since the gradient of the Lagrangian is a first order approximation of the direction in which a variable will move in the next step, we can take the sign of the entries of μ_k to estimate

whether the respective component of a_k will move into the feasible or infeasible direction in the next step, if it is at the bounds now. Active set methods then fix those parameters a_k^i at the bound a_0 if they are already at the bound and are expected to move into the infeasible direction. In order to guarantee convergence $W_k \to \mathcal{I}_a$, practical methods impose a set of additional rules on the choice of W_k in each step.

The discretized case

If we consider the discretized Newton steps, we need to define which parameter degrees of freedom we want fix in each step. The working sets W_k are again sets of indices i and can be determined by Lagrange multipliers that we will discretize in the same way as the parameter a_h itself. A semi-formal derivation yields that in analogy to (3.4), a continuous Lagrange multiplier can be defined by

$$(\mu_k, \chi)_{L^2} = -\nabla_a L(x_k, \chi) \qquad \forall \chi \in \mathcal{A}.$$

Without attempting to justify this formula in a strict sense, we discretize the Lagrange multiplier. For this purpose, recall that as basis for the parameter space $\mathcal{A}_h$ we have chosen the shape functions $\{\chi_i\}$ from $Q_d^r(\mathbb{T}_a)$. Then, $a_{k,h} = \sum_i a_{k,h}^i \chi_i$, and we likewise define a discrete Lagrange multiplier by $\mu_h = \sum_i \mu_h^i \chi_i$. With this, we define the approximate Lagrange multiplier by

$$(\mu_h, \chi_h) = -\nabla_a L(x_k, \chi_h) \qquad \forall \chi \in \mathcal{A}_h,$$

i.e. μ_h is the L^2 projection of $-\nabla_a L(x_k; \cdot)$ onto $\mathcal{A}_h$. This quantity is easily computed, as the right hand side is already available as right hand side of the Newton step, and the left hand side only involves a mass matrix. The latter is particularly simple if discontinuous elements are used. Using this multiplier estimate, we can define the discretized working set by

$$W_k = \left\{ i : \; a_{k,h}^i = a_0 \; \wedge \; \mu_h^i < 0 \right\}.$$

Remark 3.1. *Defining the Lagrange multipliers directly on the discretized Newton step instead of the continuous level would lead to worse scaling properties. This is recognized from the observation that with the definition above, we obtain for μ_h the expression*

$$\mu_h = -M_a^{-1} J^a,$$

with $(J^a)_i = \nabla_a L(x_k; \chi_i)$, while a definition of the Lagrange multipliers on the discrete set would yield a similar formula but with the mass matrix M_a on $\mathcal{A}_h$ replaced by the identity matrix. For nonuniform meshes, this results in Lagrange multipliers of which the sizes are no more comparable.

Selecting the active set

Most standard active set methods are not suited for large numbers of constraints, or infinite dimensional problems, since they allow only one constraint

to be added or removed from the working set in each step, and require that a quadratic problem is solved upon each change in the working set. This results in an exponential growth of the worst case numerical effort with the number of constraints. Although this worst case behavior rarely occurs in practice, the actual number is still prohibitively high (at least linear in the number of parameters) for the problems considered in this work.

These methods are therefore not applicable for the problems we consider here, for at least two reasons:

- We consider problems with up to several thousand parameters, each constrained by lower and upper bounds. Thus, any attempt to solve one quadratic subproblem per change in the active set is doomed to exceed computational possibilities.

- On each grid, we only make a small number of Newton steps. Since we do not aim for high accuracy on a fixed grid, there is not point in aiming at identifying the active set exactly.

Therefore, we use a modified approach where we choose the active set independently in each Newton step, and only solve one quadratic problem with this set rather than iterating with the same quadratic model until we have found the exact active set for this step. This has the drawback that we cannot prove that we do not run into a cycling active set, but it has the advantage that we can treat even very large problems. In practice, this strategy has proven successful in all applications.

The complexity of the method can be inferred from the following considerations:

- Before each step, the active set is determined by looking at those parameters that are already at their bounds, and the gradient with respect to the coefficient. Since this gradient is available for the Newton step anyway, this is cheap.

- We then fix some parameters and solve for the Newton step with these equality constrained parameters. Since fixing these parameters is equivalent to deleting the respective rows and columns of the full or reduced Hessian and the right hand side, this step is not more expensive than solving the unconstrained problem.

- Deleting rows and columns is simple even if a matrix is not known explicitly but only by application to a vector. Therefore, this approach is simple to implement also for the case of the Schur complement (reduced Hessian) method used in this thesis (see Section 1.7).

- As an iterative scheme is used to invert the Schur complement matrix, we note that deleting fixed parameters reduces the size of the matrix and the condition number of the resulting matrix is at least not larger than before. Since often a non-negligible number of parameters is fixed, the condition number may even be significantly smaller, accelerating convergence.

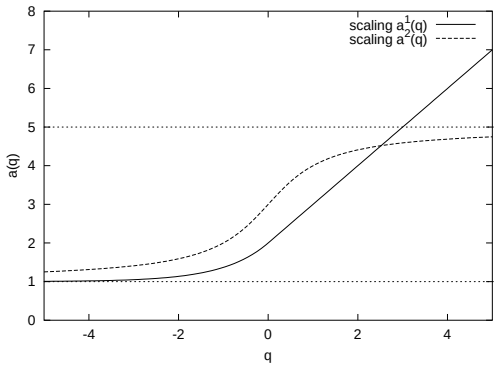


Figure 3.1: *Scaling functions for the parameter.*

Summarizing, the proposed method yields a very efficient scheme that is successfully applied even to very large problems with several thousand parameter degrees of freedom.

3.2 Treating parameter bounds by transformation

As an alternative to the active set method introduced in the previous section, we tried to handle parameter bounds by transformation. To do so, we introduce a new variable $q(\mathbf{x})$ and define a unique, strictly monotone function $a = a(q)$ such that

$$0 < \alpha \leq \inf_{q \in \mathbb{R}} a(q) \leq a_0,$$
$$a_1 \leq \sup_{q \in \mathbb{R}} a(q) \leq \infty.$$

One may choose infimum and supremum of $a(q)$ equal to a_0 and a_1, respectively, in which case the bound constraints are satisfied exactly. In practice, however, it may be better to allow for a certain violation of these bounds and only enforce $0 < \alpha \leq \inf_q a(q)$ strictly, in order to reduce the nonlinearity in the working range $a_0 \leq a \leq a_1$, and to avoid bad scaling. Possible scaling functions that were tried are

$$a^1(q) = \begin{cases} a_0(\exp(q) + 1) & \text{for } q < 0, \\ a_0(q + 2) & \text{for } q \geq 0; \end{cases} \qquad a^2(q) = \frac{2c_1}{\pi} \arctan(q) + c_2,$$

where for the second function $c_1 = \frac{1}{2}(\tilde{a}_1 - \tilde{a}_0)$, $c_2 = \frac{1}{2}(\tilde{a}_1 + \tilde{a}_0)$, $\tilde{a}_0 > 0$, and $[\tilde{a}_0, \tilde{a}_1] \supset [a_0, a_1]$ is an interval that includes (but may be larger than) the range of physical parameters. Fig. 3.1 shows a plot of these two scaling functions, with $\tilde{a}_i = a_i, i = 1, 2$. To help in the convergence of Newton steps, one should in practice use a smoother version of a^1, for example in C^2 or even C^∞.

Compared to the active set strategy, enforcing bounds by transformation did not work too well. This can, among other factors, be attributed to the increase in nonlinearity, forcing small step lengths and thus slowing down convergence.

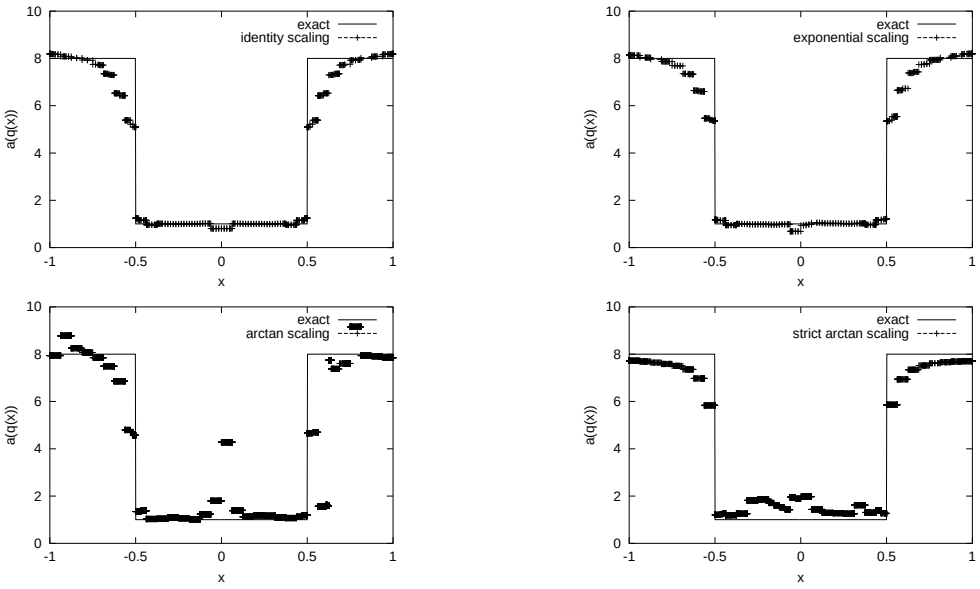


Figure 3.2: *Comparison of identified coefficient for various scaling functions.*

Furthermore, the approximation of the exact coefficient was not as good as when the active set strategy was used, again due to the fact that an exact enforcement of the bounds is only possible at the cost of strong nonlinearities.

Fig. 3.2 shows the results of a comparison for a one-dimensional version of test case 2 (see page 37). The "identity scaling" used $a(q) = q$, while the "exponential" and "arc tangent" scaling used the functions defined above with constants set such that the allowed range is slightly larger than the exact maximal and minimal values. For the "strict arc tangent" scaling, the bounds are enforced exactly. The identity scaling was included for comparison.

Although in this noise free case all scalings should theoretically recover the solution well, it is obvious that the identity scaling is the best strategy. This holds for other cases as well, unless the recovered coefficient becomes negative where the identity scaling fails, of course.

Summarizing, this approach is significantly less well suited to the problems under consideration, compared to the active set strategy. It is useful to enforce positivity of coefficients but its drawbacks prevent its use for more practical applications.

3.3 Treating parameter bounds by projection

Another, simpler, but equally unsuccessful, alternative is to compute the search direction δx_k as in the Newton step without any constraints on the bounds, but then only consider the projection onto the feasible set with respect to these bounds,

$$x_{k+1} = P_{[a_0,a_1]}\left(x_k + \alpha_k \delta x_k\right),$$

where the projector $P_{[a_0,a_1]}$ applied to $x = \{u, a, \lambda\}$ is defined by

$$P_{[a_0,a_1]}u = u, \qquad P_{[a_0,a_1]}a = \begin{cases} a_0 & \text{if } a < a_0 \\ a & \text{if } a_0 \leq a \leq a_1 \\ a_1 & \text{if } a > a_1 \end{cases}, \qquad P_{[a_0,a_1]}\lambda = \lambda.$$

Unfortunately, this approach fails since the search directions are becoming almost perpendicular to the constraint. Only back-projecting the parameters while not touching the state variable accordingly then introduces a strong violation of the state equation, which forces us to take small steps.

The solution to this is to first project the new parameter onto the feasible set, and from this compute state and adjoint variable. The drawback of this is that, again, we project away the larger part of the computed parameter update once search directions are become mostly orthogonal to the constraints. However, if we do not solve the linear equations to very high equations, the remaining small tangential component of the update is then dominated by the iteration error and is useless as a search direction.

Thus, if we do not want to solve the linear Newton steps to high accuracy, it is necessary to project away constrained parameters before, rather than after solving. This is what the active set strategy discussed above basically does.

3.4 Results

In this chapter, three methods for the incorporation of bound constraints have been discussed. While the approach using a transformation of the parameter suffers from ill-conditioning, the chosen active set strategy allows to solve even very large problems, with up to thousands of parameters, at the same or even lower numerical cost as for the unconstrained problem.

A third approach using a projection of the search direction, was shown to be related to the active set method, but suffered from problems when linear systems are not solved to high accuracies. In that case, the remaining update direction after projection includes the amplified error from the inexact iterative inversion of the matrix. In usual applications, the inversion of the Hessian is only performed up to a reduction in linear residual of 10^{-2} or 10^{-3} compared to the initial residual. This explains why the resulting search directions of the projection method are too inaccurate for practical purposes. Therefore, the projection method can only be used if a significantly higher numerical cost is accepted. Since this is hardly possible for the large scale problems discussed here, the projection method is not an option.

Finally, we briefly present one example of using bounds in the identification process. The method used was the active set strategy discussed above. We consider test case 2, where the exact coefficient varies between $a_0 = 1$ and $a_1 = 8$. Fig. 3.3 shows the identified coefficient after a number of iterations, with and without noise, and for different bounds imposed. The situation in the right column where we impose exact bounds corresponds to an identification problem where we know that the material is composed of two parts, but the interface is unknown.

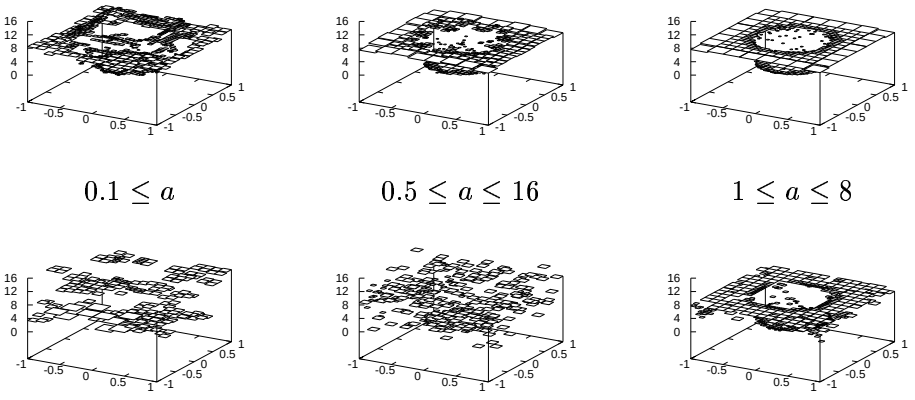


Figure 3.3: *Effect of incorporating bounds. Top row: No measurement error. Bottom row: 2% measurement error. The actual bounds imposed in each column are stated in the middle.*

It is clear that incorporating bounds acts as a stabilization, even if not the exact lower and upper bounds are chosen. In the case of no bounds (left column), the identified coefficient transgresses the shown range in some parts of the domain already for the case of no noise. On the other hand, with noise added, the identification using bounds is rather stable.

Chapter 4

Multiple experiments

In many applications we have multiple measurements z^i. For example, we may have a situation where we have a high noise level in our measurements and choose to measure several times for the same situation, or for different sources, in order to reduce the effect of the noise to the uncertainties in the recovered coefficients. Or, we may be in a situation where one measurement is not even sufficient to recover the coefficients. A similar situation arises if experiments are not set up willingly, but if naturally occurring situations are used for measurements, for example signals generated by earthquakes; we will subsume this case likewise with the term *multiple experiments*.

A similar situation is so-called *multi-physics inversion*: we try to recover parameters from different types of experiments. For example, subsurface imaging in geophysical prospection is often conducted by collecting data from entirely different sources, for example from seismic imaging, gravimetry data (recording the local gravitational force on a unit weight at different places, usually measured by flying a gravimeter over the target area), magnetotelluric data (recording the local magnetic fields), DC resistivity (measuring the electric field for a given potential), etc. These measurements are described by a set of different state equations, but depend in some way or other on the same set of parameters (density, elasticity, ...) which we would like to recover. Since each measurement alone may have little or no sensitivity for certain parameters, joint inversion is often the only possibility to obtain a set of consistent parameters.

In this chapter, we discuss a mathematical formulation for multiple experiments potentially described by multiple physics and briefly describe a framework for the implementation of such a program. Examples will be given for the case of multiple experiments for the elliptic equation considered in the previous chapters. Further applications involving the Helmholtz equation will be discussed in the final chapter.

4.1 Mathematical formulation

Multiple experiment case. Based on the statement of the problem in Chapter 1, the extension of the parameter estimation problem for measurements described by the same state equation is simple. Considering the case that the state

equation is the parameter-dependent diffusion equation discussed in previous chapters, each measurement now is characterized by a different set of applied boundary data g^i and right hand sides f^i. For the set of all these experiments, let us define the vectors $\mathbf{z} = \{z^1, \dots, z^N\}$ of measurements, $\mathbf{u} = \{u^1, \dots, u^N\}$ of solutions, and $\boldsymbol{\lambda} = \{\lambda^1, \dots, \lambda^N\}$ of Lagrange multipliers, where N denotes the number of experiments made. We assume that in all realized experiments the parameter $a(\mathbf{x})$ is unchanged. Goal is then the minimization of deviations $m(u^i - z^i)$ subject to the constraint that the $u^i \in V_{g^i}$ satisfy the state equations

$$\left(a\nabla u^i, \nabla\varphi\right) = (f^i, \varphi) \qquad \forall\varphi \in V_0.$$

Assuming equal noise levels on all measurements, i.e. giving all observations the same weight, we define the Lagrangian in analogy to Problem 1.8 to be

$$L(x) = \sum_{i=1}^{N} m(u^i - z^i) + \beta r(a) + \sum_{i=1}^{N} \left[\left(a\nabla u^i, \lambda^i\right) - (f^i, \lambda^i)\right], \tag{4.1}$$

where $x = \{\mathbf{u}, a, \boldsymbol{\lambda}\} \in \mathcal{X}_g = V_{g^1} \times \cdots \times V_{g^N} \times \mathcal{A} \times V_0 \times \cdots \times V_0$. The corresponding first order conditions for solutions $x \in \mathcal{X}_g$ then read:

$$\nabla_x L(x; y) = 0 \qquad \forall y \in \mathcal{X}_0, \tag{4.2}$$

where $\mathcal{X}_0 = V_0 \times \cdots \times V_0 \times \mathcal{A}'[a] \times V_0 \times \cdots \times V_0$ is the tangential cone to $\mathcal{X}_g$. This nonlinear system is solved using Newton's method in the same way as in Section 1.3.

Multi-physics case. Generalizing the formulation above to the case of different state equations describing the different measurements, joint inversion is described by the following quantities:

- The solutions u^i are now different quantities, having different units and meanings, each denoting the measured quantity of one experiment.
- The single coefficient a is now in general a whole set of parameters, some of them possibly space or time dependent.
- The governing equations are described by different operators $\mathcal{A}^i$ and right hand sides f^i. Not all experiments need to be sensitive to all coefficients, i.e. each $\mathcal{A}^i$ may depend on only a subset of a.
- The measurement functionals $m^i(\cdot)$ are different. For example, they may evaluate time series, scalar or spatially distributed values. Also, they may have different noise and confidence levels associated with them, which we incorporate by associating different weights σ^i to each of them.

All this is included in the following joint formulation, analogous to Problem 1.7:

Problem 4.1 (Continuous problem). *Minimize the regularized deviation*

$$J(\mathbf{u}) = \sum_{i=1}^{N} \sigma^i m^i(u^i - z^i) + \beta r(a)$$

of the u^i from the measurements z^i, with β being a regularization parameter, subject to the state equations

$$A^i(a; u^i\varphi^i) - (f^i, \varphi^i) = 0 \qquad \forall \varphi^i \in V^i,$$

where $A^i(\cdot;\cdot,\cdot)$ are the semilinear forms associated with the operators $\mathcal{A}^i$ and the set of parameters a, as well as to boundary and initial conditions and constraints on the parameters

$$a_0 \leq a \leq a_1.$$

This problem is transformed into a Lagrangian formulation in the same way as in Section 1.2.

4.2 Solution of the linear problems

After defining and discretizing the Newton step for the multiple experiment case in the same way as in Section 1.5, we are faced with the solution of the following system of linear equations in each Newton step completely analogous to the system (1.15):

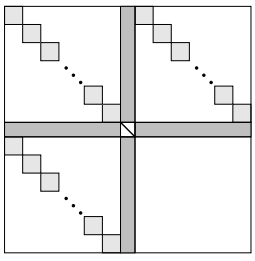
$$\begin{pmatrix} M & B^T & A^T \\ B & R & C^T \\ A & C & 0 \end{pmatrix} \begin{pmatrix} \delta \mathbf{u}_k \\ \delta a_k \\ \delta \boldsymbol{\lambda}_k \end{pmatrix} = \begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix}. \tag{4.3}$$

Matrices and vectors are now composed of blocks for the different experiments. By considering the dimension of this system,

$$n = \sum_{i=1}^{N} \#u_k^i + \#a_k + \sum_{i=1}^{N} \#\lambda_k^i,$$

where $\#\varphi$ denotes the number of degrees of freedom in the discretized variable φ, it is obvious that a direct or iterative solution of the entire system is not possible if we have more than a small number of experiments.

However, since measurements and state equations for different experiments are independent of each other and are only coupled by the common set of parameters, the system matrix above has the following block structure:



Using the Gauß-Newton modification, the Schur complement of this matrix allows the reformulation of (4.3) to the following equation for δa_k,

$$\left\{ R + \sum_{i=0}^{N} C_i^T A_i^{-T} M_i A_i^{-1} C_i \right\} \delta a_k = F_2 - \sum_{i=1}^{N} C_i^T A_i^{-T} \left(F_1^i - M_i A_i^{-1} F_3^i \right), \tag{4.4}$$

and then in turn for the single experiment state and adjoint variables δu_k and $\delta\lambda_k$

$$\delta u_k^i = A_i^{-1}(F_3^i - C_i\delta a_k), \tag{4.5}$$
$$\delta\lambda_k^i = A_i^{-T}(F_1^i - M_i\delta u_k^i). \tag{4.6}$$

Therefore, assuming we have a solver for the single operator matrices A_i, we can invert the large system with an effort that is proportional to the number of experiments N. Furthermore, since the solution of forward and adjoint equations for different experiments, as well as setting up the right hand sides is independent between experiments, the solution of the system can easily be parallelized.

4.3 Implementation

If many experiments are involved in one inversion, the numerical solution can be challenging: as each experiment requires memory resources of the same order as one forward problem, single computers can quickly become too small for a multiple experiment inversion problem. Also, since we need many nonlinear steps and many solutions of forward and backward problems are necessary in each nonlinear step, computing time requirements are even higher.

For this reason, an approach has been developed to abstract the implementation of experiments to a simple interface between a master process and slaves, each slave representing one experiment. Using this abstraction, individual experiments are sealed entities of which only the interface exists outside. While this makes the implementation of the master process more complex, it allows the simple placement of slaves on different computers, thus using the computational resources of workstation clusters.

In Fig. 4.1, the requirements on the interfaces of the three most important classes representing the master and the individual experiment slaves, as well as the description of parameters are listed. In the actual implementation, the objects need to provide a few additional functions that are used mostly for bookkeeping, such as computing the misfits or errors.

The interfaces of the different classes are strictly separated and kept minimal. Information flow between distinct modules of the program is restricted to a minimum, and different objects only communicate through their interfaces, rather than by accessing mutual data. This way, it is possible to only provide the interface on one computer while computations are actually performed on a different one, thus parallelizing the computations for different experiments. Since passing parameter vectors to functions from this interface is relatively cheap compared to the actual computations done on them, the speed-up when using multiple computers is almost optimal.

Furthermore, since the interface above is fixed, it is simple to extend the program with additional equations describing different settings; for the master process, the addition of a new experiment type is transparent.

Master object

Manages data:

- Manages slave objects
- Manages representation of the parameter
- Present value a_k of the parameters

Provides functions:

- `compute_rhs`, `compute_updates`, `perform_updates`, `compute_residual`: distribute work to respective functions of parameter representation and slaves
- `solve`:
  - *out:* solution δa of the first equation of the Schur complement system (4.4)

Parameter representation

Manages data:

- Description of parameters, e.g. discretization of a distributed field
- Description of bounds on each parameter

Provides functions:

- `compute_rhs`:
  - *out:* F_2
- `multiply_by_R/M`:
  - *in:* a
  - *out:* Ra, $M_a a$

Each slave object

Manages data:

- State mesh $\mathbb{T}^i$
- Present values u_k^i, λ_k^i of the state and adjoint variables
- Copy of present parameter a_k

Provides functions:

- `compute_rhs`:
  - *compute:* F_1^i, F_3^i
  - *out:* $C_i^T A_i^{-T}(F_1^i - M_i A_i^{-1} C_i F_3^i)$
- `forward_backward`:
  - *in:* test vector δa
  - *out:* $C_i^T A_i^{-T} M_i A_i^{-1} C_i \delta a$
- `compute_updates`:
  - *in:* update δa_k
  - *compute and store:* $\delta u_k^i = A_i^{-1}(F_3^i - C_i \delta a_k)$ $\delta \lambda_k^i = A_i^{-T}(F_1^i - M_i \delta u_k^i)$
- `perform_updates`:
  - *in:* step length α_k, δa_k
  - *compute and store:* $u_{k+1}^i = u_k^i + \alpha_k \delta u_k^i$ $\lambda_{k+1}^i = \lambda_k^i + \alpha_k \delta \lambda_k^i$
- `compute_residual`:
  - *in:* test step length α_k, δa_k
  - *out:* residual for this step length

Figure 4.1: *Description of the three basic interfaces of classes upon which the multiple experiment program is built. Since the interfaces are strictly separated, it is not important on which computer a certain object resides as long as its interface is available to callers.*

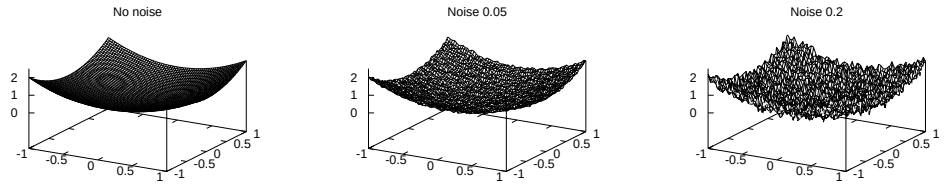


Figure 4.2: *Measurements $z(\mathbf{x})$ for different levels of added noise. Left: no noise. Center: $\varepsilon = 0.05$. Right: $\varepsilon = 0.2$.*

4.4 Application: Noise reduction

As a first example for the use of multiple experiments, we demonstrate noise reduction by multiple measurements of the same quantity. If our measurement z is subject to measurement error and other noise, then we can in general not expect to recover the exact coefficient. If we measure more than once, either with the same source or with different ones, each of these measurements will again have its uncertainties, but the coefficient that matches all the measurements best will be closer to the "correct" one because it averages over the different measurements and their errors.

In order to show the effect of measuring several times on the quality of the recovered parameter, we take test case 1 (see page 37), and put as the measurement

$$z^i(\mathbf{x}) = u^i(\mathbf{x}) + \delta^i_\varepsilon(\mathbf{x}),$$

where $\delta^i_\varepsilon(\mathbf{x})$ is a function with random normally distributed values with mean value zero and ε being the standard deviation, i.e. the noise level. The actual representation of the noise δ^i_ε is chosen differently for each measurement. Fig. 4.2 shows typical measurements for different levels of added noise. Using these measurements, we invert for the unknown parameter on a fixed, uniformly refined mesh. For this discretization, a direct calculation shows that the best L^2 approximation is $\inf_{a_h} \|a_h - a_{exact}\|_{L^2} = 0.1177....$ Throughout this section, the grid is fixed to allow for comparisons. However, the results also hold for general, possibly adapted meshes.

The left panel of Fig. 4.3 shows the results without any regularization, i.e. $\beta = 0$: as the noise level increases, the resolution of the parameter becomes increasingly worse if only one experiment is made. If multiple measurements are available, the effect of the noise is clearly suppressed. The error can be fitted well by the dependence $\|a_h - a_{exact}\| \propto (\varepsilon/\sqrt{N})^{3/4}$ which corresponds to the well-known fact that N independent measurements reduce the uncertainty by a factor of $\sqrt{N}$.

The right panel of Fig. 4.3 shows the same experiment if we choose an optimal amount of regularization (determined by experimenting). Noise is greatly suppressed with already one experiment, yet more measurements significantly improve identification of the parameter over the case of only one experiment. The error now grows as $\|a_h - a_{exact}\| \propto (\varepsilon/\sqrt{N})^{1/3}$, indicating the effect of

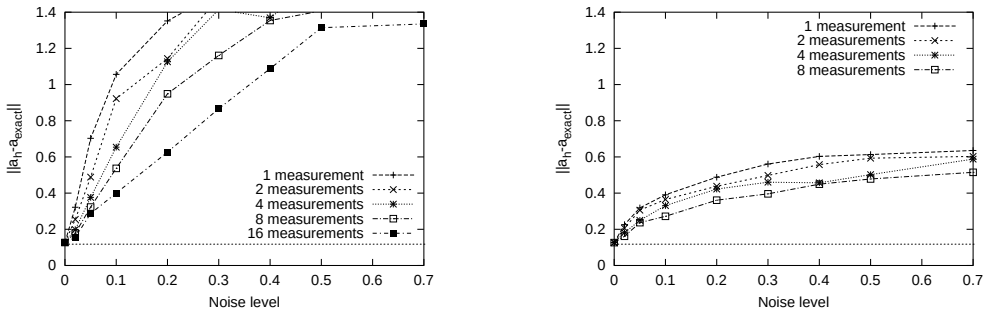


Figure 4.3: *Error* $\|a_h - a_{exact}\|$ *in the recovered coefficient for various levels of noise and numbers of measurements. Left: No regularization, i.e.* $\beta = 0$. *Right: Optimal value for* β. *The dotted line denotes the theoretical limit of approximation.*

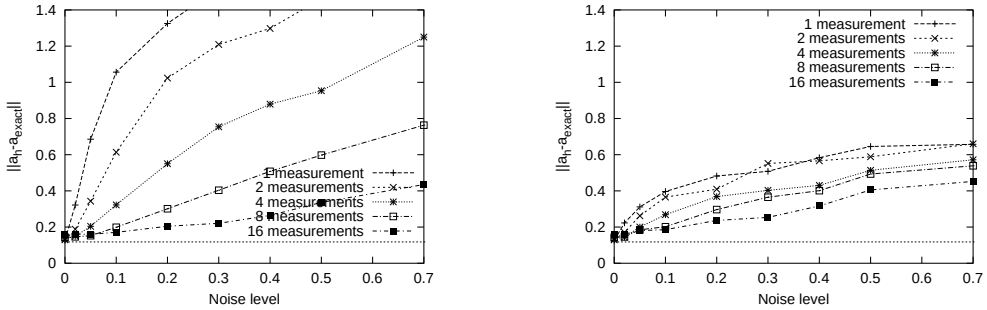


Figure 4.4: *Same as Fig. 4.3, but with different experiments using different right hand sides.*

regularization in the exponent.

The effect of noise can be even further suppressed by using different forcing functions in different experiments. Fig. 4.4 shows the results for this situation. As right hand side we use the one given in the definition of the test case (see page 37) only for the first experiment. For subsequent experiments, we use $f^i = 4\pi^2|\mathbf{k}_i|^2 \sin(2\pi\mathbf{k}_i \cdot \mathbf{x})$ with $\mathbf{k}_i \in \mathbb{N}^d$ being vectors with modulus increasing with the index i. Again, the use of several experiments can greatly improve the identification of the unknown parameter.

4.5 Application: Enforcing identifiability

Sometimes, the unknown coefficient is not identifiable at some points without regularization. For example, in the one-dimensional case, assuming no noise, the parameter identification problem reads: *find* $a(x)$ *such that*

$$u = z, \qquad -\left(au'\right)' = f.$$

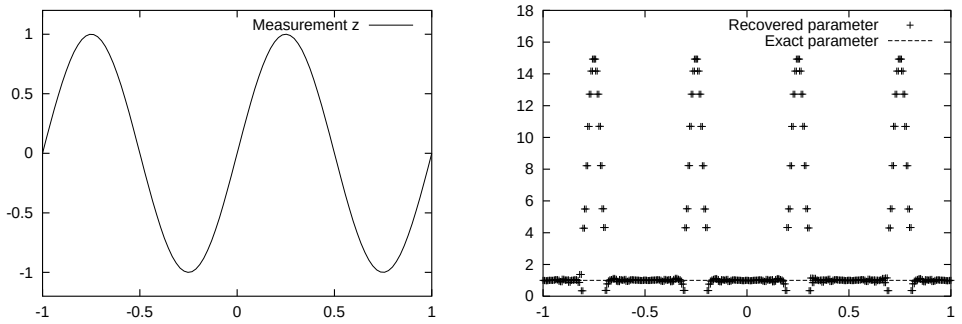


Figure 4.5: *Example of non-identifiable coefficient in one space dimension. Left: Exact displacement and measurement* $u = z = \sin(2\pi x)$. *Right: Recovered coefficient, without imposition of bounds.*

Inserting the first identity into the second yields a first order differential equation for the coefficient with analytical solution

$$a(x) = -\frac{1}{z'(x)} \int_{x_0}^{x} f(\xi) d\xi + a(x_0), \tag{4.7}$$

where x_0 denotes the left end of the interval Ω, and where $a(x_0)$ must be specified in advance. It is obvious that $a(x)$ is not identifiable at places where $z' = 0$. Likewise, the coefficient is not identifiable in higher dimensions at places where $\nabla z = 0$, although the proof of ill-posedness there is more difficult (see, for example, Banks and Kunisch [13]).

This concept of identifiability only concerns single points. For L^∞ coefficients, we could simply ignore such points. However, the coefficient is usually badly resolved also in their environment, spoiling the identification process. Fig. 4.5 shows this in one space dimension. We choose $u = \sin(2\pi x)$, no noise (i.e. $z = u$), $a = 1$ and thus $f = -u''$. We do not use regularization and do not impose bounds on the coefficient. The recovery of the coefficient is clearly insufficient near points where $u' = 0$.

Adding regularization to the minimization problem allows to identify a coefficient although it is solely determined by the regularization at points where $z' = 0$, not by measurement. The left panel of Fig. 4.6 shows the result for an optimal amount of regularization. After the last iteration, the error is $\|a_h - a_{exact}\| = 0.17$.

Instead, we can also perform several experiments in such a way that at no point all measurements have $(u^i)' = 0$. For example, we might choose the forces f^i such that $u^1 = \sin(2\pi x)$ and $u^2 = \sin(3\pi x)$. The result is shown in Fig. 4.6. The error in the coefficient after the last iteration is now $\|a_h - a_{exact}\| = 0.00013$, i.e. approximately a factor of 1000 smaller than the result obtained with regularization.

The importance of this lies in the fact that for some setups of physical experiments, whole regions are unidentifiable. For example in an imaging experiment, entire regions may lie in the shadow. Then, several experiments

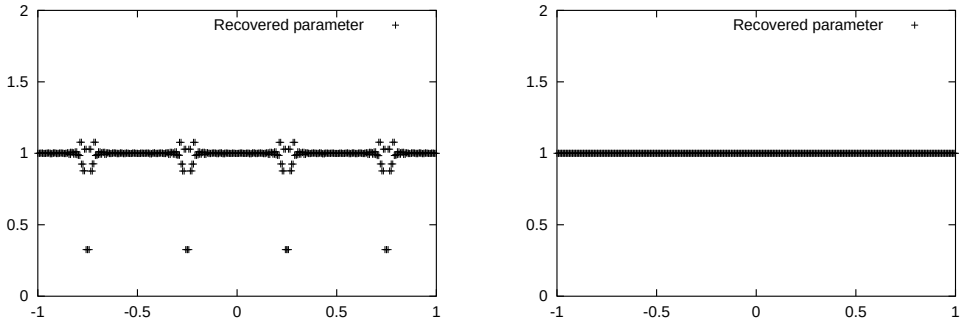


Figure 4.6: *Example of non-identifiable coefficient in one space dimension. Left: Recovered coefficient with optimal regularization and one experiment. Right: Recovered coefficient without regularization and two experiments. Note the different maximal errors compared to Fig. 4.5.*

illuminating from different angles may help to identify the solution. Multiple measurements with different sources are therefore commonly used in seismic imaging experiments.

Chapter 5

Inverse wave problems

In this chapter, we will apply the techniques previously developed for the diffusion equation to parameter identification problems for the Helmholtz equation. Since this is the frequency domain version of the time dependent wave equation, this class of problems is used in many applications where time dependent data is measured, for example seismic data in geophysics.

We will, in this chapter, first derive the complex valued Helmholtz equation and boundary conditions that describe the problems we consider here. Based on this, the inverse problem is formulated, and a brief comparison of the solution of inverse problems for wave problems in the time and frequency domains is given, to set a background for the methods we use here.

In the then following two sections, we briefly discuss the differences between inverse problems for the Helmholtz and the diffusion equation, then touch the two main mathematical obstacles for inverse wave problems, nonlinearity and non-uniqueness. Next, the error estimates derived in Chapter 2 are adapted to the present situation.

Finally, applications are given, that illustrate the general coefficient resolving properties of the discussed methods. Furthermore, the superior performance of weighted error estimator driven refinement over more ad hoc approaches is shown, and the accuracy of error estimates is discussed.

5.1 Inversion in frequency space

In order to state the inverse problem to be treated in this chapter in a concise way, we first define the forward problem in this section, and based on this derive the structure of the inverse problem. Although we consider a structurally time dependent problem, we formulate it in the frequency domain as a family of Helmholtz equations. The reasons for this and the resulting advantages particular to inverse problems will be discussed in a final subsection.

Formulation of the forward problem

In order to derive the equations describing the forward problem, we start with the time dependent wave equation, transfer it to frequency space by applying

a Fourier transform, and finally write it in weak form.

As starting point, we choose the time dependent wave equation on a bounded domain Ω and in a time interval $I = (0, T)$,

$$\partial_t^2 u - \nabla\cdot(a\nabla u) = 0, \qquad (\mathbf{x}, t) \in \Omega \times I, \tag{5.1}$$

with Neumann, Dirichlet, and simple absorbing boundary conditions on portions Γ_N, Γ_D, and Γ_A of the boundary $\partial\Omega$, respectively:

$$\mathbf{n}\cdot a\nabla u = 0 \qquad (\mathbf{x}, t) \in \Gamma_N \times I, \Gamma_N \subset \partial\Omega, \tag{5.2}$$

$$u = g \qquad (\mathbf{x}, t) \in \Gamma_D \times I, \Gamma_D \subset \partial\Omega, \tag{5.3}$$

$$\mathbf{n}\cdot a\nabla u + \sqrt{a}\partial_t u = 0 \qquad (\mathbf{x}, t) \in \Gamma_A \times I, \Gamma_A \subset \partial\Omega. \tag{5.4}$$

The absorbing boundary conditions chosen here are those of Bayliss and Turkell [14], which are equivalent to those of Engquist and Majda [33]. Note that these boundary conditions make the spectrum complex valued and in general continuous, even on bounded domains.

Since here we are only interested in identification of elastic properties, we have assumed that the density usually appearing before the term $\partial_t^2 u$ in (5.1) is constant. We can then scale it out of the equations. Thus, the coefficient a has the interpretation of the square of a wave speed.

We seek the solution of this set of equations in frequency space by introducing the Fourier transform u_ω of the solution u as

$$u(\mathbf{x}, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\omega t} u_\omega(\mathbf{x})\, d\omega.$$

Likewise, we define the Fourier transform g_ω of g. Inserting these functions into equations (5.1)–(5.4) then yields

$$-\omega^2 u_\omega - \nabla\cdot(a\nabla u_\omega) = 0 \qquad \mathbf{x} \in \Omega, \tag{5.5}$$

$$\mathbf{n}\cdot a\nabla u_\omega = 0 \qquad \mathbf{x} \in \Gamma_N, \tag{5.6}$$

$$u_\omega = g_\omega \qquad \mathbf{x} \in \Gamma_D, \tag{5.7}$$

$$\mathbf{n}\cdot a\nabla u_\omega + i\omega\sqrt{a}u_\omega = 0 \qquad \mathbf{x} \in \Gamma_A. \tag{5.8}$$

These equations have to be solved for each member u_ω of a family indexed by $\omega \in \mathbb{R}$.

The problem is transformed into a weak formulation in the usual way. It then reads: *find* $u_\omega \in V_{g_\omega} = \{u_\omega \in H^1(\Omega \to \mathbb{C}) : u_\omega|_{\Gamma_D} = g_\omega\}$, *such that for all* $\varphi \in V_0 = \{\varphi \in H^1(\Omega \to \mathbb{C}) : \varphi|_{\Gamma_D} = 0\}$ *there holds*

$$-(\omega^2 u_\omega, \varphi)_\Omega + (a\nabla u_\omega, \nabla\varphi)_\Omega + (i\omega\sqrt{a}u_\omega, \varphi)_{\Gamma_A} = 0,$$

again for every frequency ω. Splitting this equation into its real and imaginary parts, and denoting $u_\omega = v_\omega + iw_\omega$, we obtain the final form of the state equation for each component in ω-space:

Problem 5.1 (Forward problem). *For each $\omega \in \mathbb{R}$, find the solution $u_\omega = \{v_\omega, w_\omega\} \in V_{g_\omega} = \{u_\omega \in H^1(\Omega \to \mathbb{R})^2 : v_\omega + iw_\omega|_{\Gamma_D} = g_\omega\}$, such that for all $\varphi = \{\zeta, \xi\} \in V_0 = \{\varphi \in H^1(\Omega \to \mathbb{R})^2 : \varphi|_{\Gamma_D} = 0\}$ there holds*

$$A_\omega(a; u_\omega, \varphi) = 0, \tag{5.9}$$

with the bilinear form

$$\begin{aligned} A_\omega(a; u_\omega, \varphi) = &-(\omega^2 v_\omega, \zeta)_\Omega + (a\nabla v_\omega, \nabla\zeta)_\Omega - (\omega\sqrt{a} w_\omega, \zeta)_{\Gamma_A} \\ &+ (\omega^2 w_\omega, \xi)_\Omega - (a\nabla w_\omega, \nabla\xi)_\Omega - (\omega\sqrt{a} v_\omega, \xi)_{\Gamma_A}. \end{aligned}$$

Note that we have deliberately reversed the sign of the equation defining the imaginary part, making A_ω symmetric, i.e. $A_\omega(a; u_\omega, \varphi) = A_\omega(a; \varphi, u_\omega)$. This has positive effects on the solvability of the discretized equations using iterative schemes.

The inverse problem

The inverse problem of estimating the distributed parameter a in (5.9) is formulated similar to the one discussed throughout previous chapters. Adopting the same notation regarding misfit and regularization functionals $m(\cdot)$ and $r(\cdot)$, the inverse problem in the single experiment case reads in analogy to Problem 1.7:

Problem 5.2. *Minimize the regularized deviation*

$$J(u, a) = m(u_\omega - z_\omega) + \beta r(a)$$

of $u_\omega = \{v_\omega, w_\omega\}$ from the measurement z, with β being a regularization parameter, subject to the state equation (5.9), and the additional constraints

$$\begin{aligned} u_\omega|_{\Gamma_D} &= g_\omega, \\ a_0 \le a &\le a_1. \end{aligned}$$

The characterization of the solution by a Lagrange functional and its stationary points then follow in the same way as in Problem 1.8.

In general, one is not interested in inverting only one measurement with only one frequency component. In this case, the misfit functional $m(\cdot)$ will contain a sum over those frequencies ω_i for which measurements exist. The different frequency components u_{ω_i} then each have to satisfy a state equation with semilinear forms A_{ω_i}, resulting in a multiple experiment situation as discussed in Chapter 4. We will only consider this multiple experiment case in the following. The corresponding formulation of the identification problem then is as in Problem 4.1. We explicitly show it here for later reference:

Problem 5.3. *Solutions of the multiple experiment Helmholtz inversion problem are characterized by stationary points*

$$\nabla_x L(x; y) = 0 \qquad \forall y \in \mathcal{X}_0,$$

of the Lagrangian $L(x)$, where $x = \{u_{\omega_1}, \ldots, u_{\omega_N}, a, \lambda_{\omega_1}, \ldots, \lambda_{\omega_N}\} \in \mathcal{X}_g$, $\mathcal{X}_g = (\prod_{i=1}^N V_{g_{\omega_i}}) \times \mathcal{A} \times V_0^N$, $\mathcal{X}_0 = V_0^N \times \mathcal{A} \times V_0^N$, and V_{g_ω}, V_0 as defined in Problem 5.1. The Lagrangian is defined by

$$L(x) = J(x) + \sum_{i=1}^{N} A_{\omega_i}(a; u_{\omega_i}, \lambda_{\omega_i}), \tag{5.10}$$

with the form A_ω as in Problem 5.1, and

$$J(x) = \sum_{i=1}^{N} m(u_{\omega_i} - z_{\omega_i}) + \beta r(a).$$

Applying Newton's method to the optimality condition, and discretizing each step then leads to a system of linear equations equivalent to (4.4)–(4.6), where we now have the matrices

$$M_i = \begin{bmatrix} M & 0 \\ 0 & M \end{bmatrix}, \quad A_i = \begin{bmatrix} A_{\omega_i} & -G_{\omega_i} \\ -G_{\omega_i} & -A_{\omega_i} \end{bmatrix}, \quad C_i = \begin{bmatrix} C_1(v_{\omega_i}) - \omega_i C_2(w_{\omega_i}) \\ -C_1(w_{\omega_i}) - \omega_i C_2(v_{\omega_i}) \end{bmatrix},$$

composed of the following blocks:

$$M_{kl} = (m^i)''(u_{\omega_i}; \varphi_k, \varphi_l), \qquad A_{\omega_i,kl} = -(\omega_i^2 \varphi_l, \varphi_l)_\Omega + (a \nabla \varphi_l, \nabla \varphi_l)_\Omega,$$
$$G_{\omega_i,kl} = (\omega_i \sqrt{a} \varphi_i, \varphi_j)_{\Gamma_A},$$
$$C_1(p)_{kl} = (\nabla p \cdot \nabla \varphi_k, \chi_l)_\Omega, \qquad C_2(p)_{kl} = \left(\frac{1}{2\sqrt{a}}\, p\, \varphi_k, \chi_l \right)_{\Gamma_A},$$

with φ_k, χ_l being the trial functions for primal and dual variables, and parameter variables, respectively.

Comparison between time and frequency domain

Above, we have used the frequency domain to formulate the problem of identification of parameters in a time dependent wave equation. While conceptually solving in the time or the frequency domain is equivalent, there are significant differences when numerically approximating the forward solution on a computer:

- In the time domain, a time-stepping scheme is used to solve the subproblems on subsequent time steps. This generates the sought solution directly, but each time step depends on the prior solution of the last time step. In each time step, the solutions of at least two time steps have to be kept in memory. The number of time steps is roughly proportional to the highest frequency occurring.

- In the frequency domain, the solutions u_{ω_i} for different frequencies ω_i do not depend on each other. This allows for simple parallelization. However, if we are interested in the wave field in the time domain, this can only be computed by overlaying the solutions of *all* components and forming the Fourier back-transform of it. The number of frequency components that

have to be computed for a given accuracy of the time dependent wave field is proportional to the size of the frequency band that is excited by sources.

For the present application, solution in the frequency domain is more adequate for three reasons:

- We are not interested in the time dependent wave field, but only in the comparison to given measurements. This can be done in the time and frequency domains equally well.
- In applications, often only small frequency bands are excited. While in the time domain, the numerical effort is proportional to the *highest* frequency, in the frequency domain it is only proportional to the *size* of the frequency band. As an extreme case, consider time harmonic excitations: we would then only have to solve one problem in the frequency domain, but still many time steps in the time domain.
- Since the problems in frequency domain are independent of each other, we can use this to parallelize the problem in the same way as described in Chapter 4.

As will be explained in Section 5.3, an additional reason for inverting in the frequency domain is the stabilization of the problem if one starts with low frequencies, as this reduces the nonlinearity.

5.2 Comparison with diffusion problems

Compared to the static problems governed by a diffusion equation discussed in the previous chapters, the problems considered here differ in several respects concerning computational complexity. In this section, we briefly review why the problems of this chapter are more challenging. A discussion of mathematical problems arising with typical inverse problems for the Helmholtz equation is given in the next section.

The foremost reason for the higher complexity is that solving the Helmholtz equation numerically is more difficult than the Laplace equation. This has three main reasons:

- The indefiniteness of the operator disallows the use of simple conjugate gradient methods. If ω^2 is too close to an eigenvalue of the Laplace operator, the problem is also ill-conditioned.
- The traveling wave character of solutions of the Helmholtz equation results in solutions that do not decay quickly with the distance to sources, requiring mesh refinement in large parts of the domain.
- Solutions of the Helmholtz equation are oscillatory, where the wavelength of solutions is $\lambda \propto 1/\omega$. This requires fine meshes especially for high frequencies. For good resolution, the mesh width should satisfy at least $\lambda/h \geq 10$.

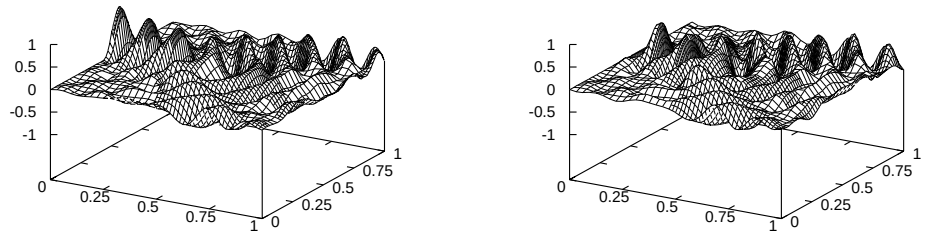


Figure 5.1: *Typical solution $u_\omega = v_\omega + iw_\omega$ of Helmholtz equation with $\omega = 50$. Left: Real part. Right: Imaginary part. The solution is the same as in Fig. 5.4.*

A typical solution of the state equation showing the last two points is displayed in Fig. 5.1. For this example, there are absorbing boundary conditions at the bottom, right, and top boundary, and waves are injected at the center of the left boundary.

From these considerations, it is clear that solving the Helmholtz equation is more expensive from a numerical point of view than the Laplace equation. In particular, in d space dimensions, the effort grows with the frequency ω as ω^d since the mesh width must be proportional to the wave length. For typical applications, diam $\Omega/\lambda \approx 10 \dots 100$, requiring a number of cells in the range of at least $100^d \dots 1000^d$. Finally, we remark that solving on an insufficiently fine grids leads to unusable solutions since the dispersion of finite elements results in a phase shift between exact and numerical solution. It is thus often not even possible to start on a relatively coarse mesh.

Further aspects to be taken into consideration when comparing inversion for wave and diffusion problems is that for the former we often only have boundary measurements, but in a multiple experiment setting. The fact that measurements are only on the boundary requires us to solve to relatively high accuracies. Both aspects further increase the numerical effort.

As a final comparison between wave and diffusion problem, we consider the condition numbers of Schur complement matrices. Fig. 5.1 shows the dependence on the mesh width for the Helmholtz equation with $\omega = 10$, for an otherwise comparable configuration as Table 1.2 (page 31) shows for the diffusion equation. While the growth of the condition number as the mesh is refined follows the same orders as for the diffusion equation, their absolute size is smaller, at least for the more important case of L^2 measurements, making the Newton steps simpler to solve.

On the other hand, these condition numbers also depend on the frequency, and on the number of experiments performed. Generally, the condition number decreases with higher frequencies and more experiments, making up for part of the otherwise higher complexity.

	$m(u-z)=\frac{1}{2}\|u-z\|_\Omega^2$			$m(u-z)=\frac{1}{2}\|\nabla(u-z)\|_\Omega^2$		
h	$\min\|\mu_i\|$	$\max\|\mu_i\|$	κ_2	$\min\|\mu_i\|$	$\max\|\mu_i\|$	κ_2
2^{-3}	0.00195	0.339	170	0.717	32.2	45
2^{-4}	$9.76\cdot10^{-5}$	0.160	1600	0.0916	14.8	160
2^{-5}	$2.05\cdot10^{-6}$	0.0499	24000	0.00657	4.49	680
2^{-6}	$3.97\cdot10^{-8}$	0.013	$3.5\cdot10^{5}$	0.000479	1.20	2500
2^{-7}	$7.09\cdot10^{-10}$	0.0035	$4.9\cdot10^{6}$	$3.73\cdot10^{-5}$	0.324	$8.7\cdot10^{3}$
	$\mathcal{O}(h^6)$	$\mathcal{O}(h^2)$	$\mathcal{O}(h^{-4})$	$\mathcal{O}(h^4)$	$\mathcal{O}(h^2)$	$\mathcal{O}(h^{-2})$

Table 5.1: *Minimal and maximal eigenvalues μ_i, and condition number with respect to the spectral norm for the Schur complement.*

5.3 Complications of tomography

In the applications discussed in this chapter, we only consider cases where the sources of the Helmholtz equation are located on part of the boundary, as this is typical for applications. If measurements are also performed only on the boundary, then this mode is called *tomography*. Depending on whether measurements are made at the same part of the boundary where sources are located, or on an opposite part, this is called *reflection tomography* or *transmission tomography* in the context of wave problems.

For the Laplace equation, the main problem of tomography is an extreme ill-posedness in the interior of the domain, since information entering at the boundary of the domain decays quickly as a function of the distance to the boundary. For the Helmholtz equation, just as for inversion in the time domain, this ill-posedness away from the boundary does not exist, since the corresponding Green's function has different decay properties. Nevertheless, inverting wave signals poses a number of mathematical peculiarities. Among these are strong nonlinearities of the objective function as well as non-identifiability in certain function spaces. We will briefly discuss these difficulties in this section to illustrate the typical complications of inversion for wave problems.

Nonlinearity of the inverse problem

In contrast to the diffusion problem covered in previous chapters, the objective functional usually has many local minima for the Helmholtz equation, and getting stuck in one of them is simple. To illustrate this, consider the following one dimensional example: assume we have a string of length L with constant but unknown wave propagation velocity c which we would like to recover within the range $c \in [c_0, c_1]$. We excite the string at the left end at $x = 0$ with a time-periodic signal with frequency ω and amplitude and phase $\varphi(\omega) \in \mathbb{C}$. We measure the displacement and its phase at $x = L$, where we assume that an absorbing end is placed. The frequency is chosen such that the wave length is small compared to L, i.e. $\omega \gg L/2\pi c$.

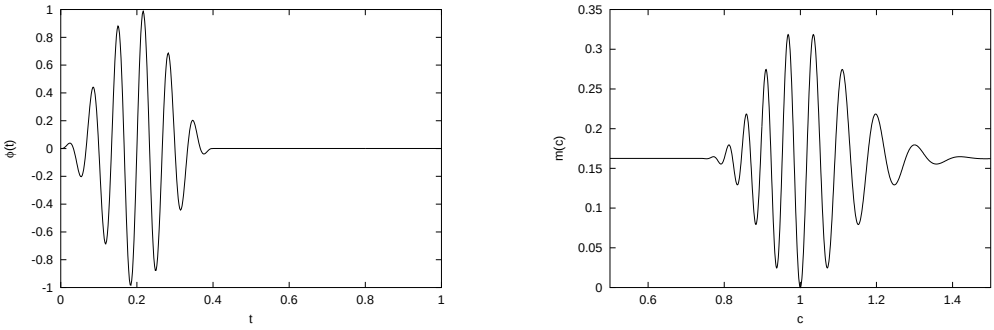


Figure 5.2: *Visualization of the nonlinearity occurring in inverse problems for wave equations. Left: Time dependent source at the left end. Right: Misfit functional $m(c)$ as function of the wave speed c.*

This situation can be described by the following set of equations:

$$-\omega^2 u_\omega - c^2 \partial_x^2 u_\omega = 0, \qquad u_\omega(0) = \varphi(\omega), \qquad (i\omega + c\partial_x) u_\omega(L) = 0. \tag{5.11}$$

The last boundary condition represents perfectly absorbing boundary conditions at $x = L$. The solution of this problem is $u_\omega(x) = \varphi(\omega)\cos(\omega x/c)$.

In the inverse problem, we are given a measurement z_ω of $u_\omega(L)$. In the noise free case, $z_\omega = \varphi(\omega)\cos(\omega L/c^*)$ with the "true" wave speed c^*. We then seek to minimize the misfit integrated over all frequencies,

$$\begin{aligned} m(c) = m(u(L) - z) &= \frac{1}{2}\int |u_\omega(L) - z_\omega|^2 \, d\omega \\ &= \frac{1}{2}\int |\varphi(\omega)(\cos(\omega L/c) - \cos(\omega L/c^*))|^2 \, d\omega \end{aligned}$$

on the range of admissible wave speeds $c \in [c_0, c_1]$. Note that by definition of the Fourier transform, the misfit in the time and the frequency domain are equivalent: $\int |u_\omega(L) - z_\omega|^2 \, d\omega = \int |u(L,t) - z(t)|^2 \, dt$.

While the solution of this problem is obvious, the misfit functional is strongly nonlinear. For example, assume we use the signal $f(t)$ shown in the left panel of Fig. 5.2. Accordingly $\varphi(\omega)$ is the Fourier transform of this signal. With $c^* = 1$, the misfit functional $m(c)$ is shown in the right panel of Fig. 5.2.

The nonlinearity of the objective functional is striking. At the center, the oscillations result from measurement and simulation being shifted relatively to each other by a fixed number of periods in the time domain and likewise by 2π in the frequency domain; small variations then bring the two functions out of phase, yielding a higher value of the misfit functional, until the wave speed changes so much that minima and maxima match once again. This phenomenon is commonly referred to as *aliasing*, since oscillations of the simulated solution match, i.e. alias, the wrong oscillations of the measurement.

The distance between two local minima, and so also the domain of attraction of a minimum, corresponds to the size of changes in the coefficient necessary to displace measurement and predicted solution by one wavelength. Thus, it is

larger for low frequencies. In practice, low frequency measurements are therefore often used to obtain a good initial guess, which is then used to proceed with high frequencies.

This nonlinearity is mostly generic for wave problems, and also exists in higher spatial dimensions. In applications, this usually leads to solutions being trapped in local minima, unless the parameter identification process is started in the close vicinity of the true solution. In applied geophysical inversion, many techniques have been developed to either generate good initial guesses, or for global optimization. The amount of literature on this is so vast that we do not attempt to give an overview. As we do not endeavor to develop techniques in this area, we will always assume that we have starting values close enough to find the desired optimum with local search techniques such as the Gauß-Newton method.

Non-uniqueness of solutions

Another difficult aspect of waveform inversion is that smooth variations of the velocity, often called the *background velocity*, are hard to determine. In fact, in one space dimension, it is not identifiable at all: let c^* be the optimal spatially constant wave speed, then in the example of the previous section all spatially varying wave speeds $c(x) = c^* + \tilde{c}$ with smooth functions $\tilde{c}$ with zero mean value will generate the same measurements at $x = L$. The reason, of course, is that we only measure the arrival times of signals, not whether it traveled faster or slower on parts of the string. The same holds, if c^* is not constant, but piecewise constant: what we see is only the arrival times of transmitted and reflected signals; these signals only contain the position of discontinuities (via the arrival times) and the height of the jumps (through the reflection amplitudes), but not the smooth variations between the jumps.

The situation is better in more than one space dimension, since there smooth variations refract waves, i.e. wave directions are bent smoothly by the background velocity, but in general the problem of determining the smooth variations is significantly more ill posed than that of recovering discontinuities.

Conclusions for examples

Since the goal of this work is not to develop techniques to work around the two problems mentioned above, we choose the examples of this chapter such that

- the sought coefficients are piecewise constant, and
- initial values are close to the exact values, but constant; therefore, the initial values do not contain a priori knowledge about the positions of jumps in the coefficient.

Both assumptions are often practicable in geophysical applications, as media in the underground are usually stratified, i.e. piecewise constant. Furthermore, good initial guesses can, for example, be obtained by traveltime inversion, which only uses the time a signal arrives, but not its amplitude and phase, thus avoiding the nonlinearity problem.

5.4 Error estimation

In this section, we briefly state the form of error estimates for the problem considered in this chapter. Since the general form of estimates in terms of the Lagrangian has already been given in Chapter 2, we only show this for estimates with respect to the minimization functional $J(\cdot)$. The form of the estimates for arbitrary functionals and for the bound constrained case can then easily be derived from this and the material of Chapter 2.

Theorem 5.4. *For the multiple experiment Helmholtz inversion problem 5.3, the error with respect to the functional $J(\cdot)$ can be represented by*

$$J(x) - J(x_h) = \frac{1}{2}\sum_{i=1}^{N}\sum_{K\in\mathbb{T}_i}\left(\eta_v^{i,K} + \eta_w^{i,K} + \eta_\zeta^{i,K} + \eta_\xi^{i,K}\right) + \frac{1}{2}\sum_{K_a\in\mathbb{T}_a}\eta_a^{K_a} + R, \tag{5.12}$$

with terms relating to the residuals of the state equation,

$$\begin{aligned}
\eta_v^{i,K} &= \left(-\omega_i^2 v_{\omega_i} - \nabla\cdot(a_h\nabla v_{\omega_i}), \zeta_{\omega_i} - i_h\zeta_{\omega_i}\right)_K \\
&\quad + \tfrac{1}{2}\left(\mathbf{n}\cdot[a_h\nabla v_{\omega_i}], \zeta_{\omega_i} - i_h\zeta_{\omega_i}\right)_{\partial K\setminus\partial\Omega} \\
&\quad + \left(\mathbf{n}\cdot a_h\nabla v_{\omega_i} - \omega_i\sqrt{a_h}w_{\omega_i,h}, \zeta_{\omega_i} - i_h\zeta_{\omega_i}\right)_{\partial K\cap\Gamma_A}, \\
\eta_w^{i,K} &= -\left(-\omega_i^2 w_{\omega_i} - \nabla\cdot(a_h\nabla w_{\omega_i}), \xi_{\omega_i} - i_h\xi_{\omega_i}\right)_K \\
&\quad - \tfrac{1}{2}\left(\mathbf{n}\cdot[a_h\nabla w_{\omega_i}], \xi_{\omega_i} - i_h\xi_{\omega_i}\right)_{\partial K\setminus\partial\Omega} \\
&\quad + \left(-\mathbf{n}\cdot a_h\nabla w_{\omega_i} - \omega_i\sqrt{a_h}v_{\omega_i,h}, \xi_{\omega_i} - i_h\xi_{\omega_i}\right)_{\partial K\cap\Gamma_A},
\end{aligned}$$

terms relating to the adjoint equation

$$\begin{aligned}
\eta_\zeta^{i,K} &= \left(v_{\omega_i,h} - Re\ z_{\omega_i} + (-\omega_i^2\zeta_{\omega_i,h} - \nabla\cdot(a_h\nabla\zeta_{\omega_i,h})), v_{\omega_i} - i_h v_{\omega_i}\right)_K \\
&\quad + \tfrac{1}{2}\left(\mathbf{n}\cdot[a_h\nabla\zeta_{\omega_i,h}], v_{\omega_i} - i_h v_{\omega_i}\right)_{\partial K\setminus\partial\Omega} \\
&\quad + \left(\mathbf{n}\cdot a_h\nabla\zeta_{\omega_i,h} - \omega_i\sqrt{a_h}\xi_{\omega_i,h}, v_{\omega_i} - i_h v_{\omega_i}\right)_{\partial K\cap\Gamma_A}, \\
\eta_\xi^{i,K} &= \left(w_{\omega_i,h} - Im\ z_{\omega_i} - (-\omega_i^2\xi_{\omega_i,h} - \nabla\cdot(a_h\nabla\xi_{\omega_i,h})), w_{\omega_i} - i_h w_{\omega_i}\right)_K \\
&\quad - \tfrac{1}{2}\left(\mathbf{n}\cdot[a_h\nabla\xi_{\omega_i,h}], w_{\omega_i} - i_h w_{\omega_i}\right)_{\partial K\setminus\partial\Omega} \\
&\quad + \left(-\mathbf{n}\cdot a_h\nabla\xi_{\omega_i,h} - \omega_i\sqrt{a_h}\zeta_{\omega_i,h}, w_{\omega_i} - i_h w_{\omega_i}\right)_{\partial K\cap\Gamma_A},
\end{aligned}$$

and finally terms involving the control equation

$$\begin{aligned}
\eta_a^{K_a} &= \left(\beta a_h + \sum_{i=1}^{N}\left(\nabla\zeta_{\omega_i,h}\cdot\nabla v_{\omega_i,h} - \nabla\xi_{\omega_i,h}\cdot\nabla w_{\omega_i,h}\right), a - i_h a\right)_{K_a} \\
&\quad - \left(\sum_{i=1}^{N}\frac{\omega_i}{2\sqrt{a_h}}(w_{\omega_i,h}\zeta_{\omega_i,h} + v_{\omega_i,h}\xi_{\omega_i,h}), a - i_h a\right)_{\partial K_a\cap\Gamma_A}.
\end{aligned}$$

The remainder term is

$$R = \tfrac{1}{2}\int_0^1 \nabla_x^3 L(x_h + se; e, e, e)\ s(s-1)\ ds.$$

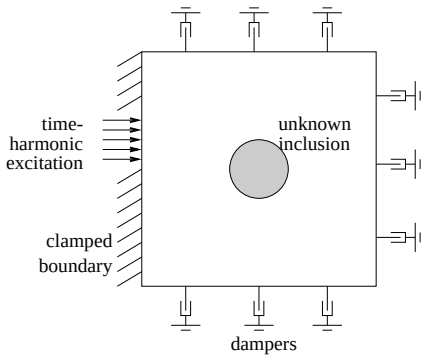


Figure 5.3: *Identification of an inclusion. Layout of example.*

Here, x and x_h are continuous and discrete solutions, respectively, and i_h is a generic interpolation operator acting on $\mathcal{X} \to \mathcal{X}_h$ or single components, depending on context.

Proof. Use the general form of the estimate in terms of the Lagrangian, given in Theorem 2.1, expand the Lagrangian (5.10), and integrate by parts on each cell. □

We will check the accuracy of this formula with the applications at the end of this chapter. Note that the other error representation formulae derived in Chapter 2 have similar forms.

5.5 Application: Identification of an inclusion

As a first example of parameter identification for the Helmholtz equation, consider the situation depicted in Fig. 5.3: a plate of elastic material is clamped at its left side and placed in dampers absorbing all waves at all other faces. A time periodic force is applied at portions of the clamped side. The position and frequency of the excitation is varied in different experiments. Finally, we assume that amplitude and phase of the resulting periodic motion of the plate can be measured at all positions; such measurements are possible with lasers, for example. The goal is to recover an unknown inclusion in the material by inverting for the spatially varying coefficient $a(\mathbf{x})$.

Given this setup, the problem can be described as follows: let the index $1 \leq i \leq N$ denote the number of the experiment, then $\mathbf{u} = \{u_{\omega_1}, \ldots, u_{\omega_N}\}, u_{\omega_i} \in V_{g_\omega^i}$ are the solutions of the state equations

$$A_{\omega_i}(a; u_{\omega_i}, \varphi) = 0 \qquad \forall \varphi \in V_0,$$

subject to boundary conditions $u_{\omega_i}|_{\Gamma_D} = g_\omega^i$, see Problem 5.1. With this constraint, the minimization problem reads

$$\min J(\mathbf{u}, a) = \sum_{i=1}^{N} \frac{1}{2} \|u_{\omega_i} - z_i\|_\Omega^2 + \frac{\beta}{2} r(a),$$

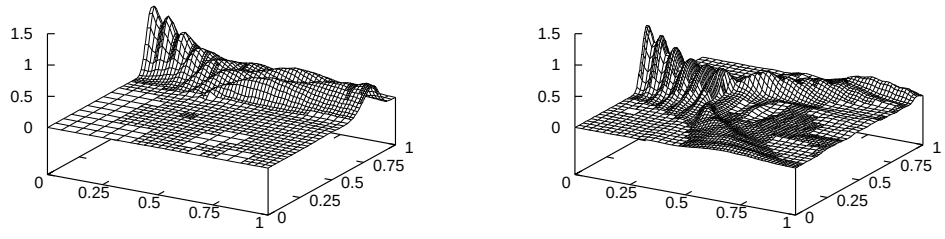


Figure 5.4: *Identification of an inclusion.* $|u_\omega|^2 = |v_\omega + iw_\omega|^2$ *for two solutions with different source positions. In both cases,* $\omega = 50$.

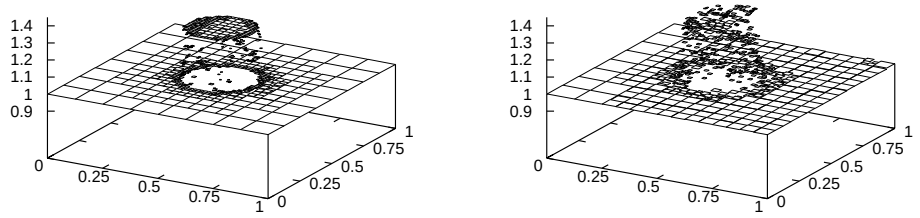


Figure 5.5: *Identification of an inclusion. Left: Identified coefficient with bounds* $1 \leq a \leq 1.3$. *Right: With bounds* $0.5 \leq a \leq 5$.

where z_i is the measurement for the ith experiment. For two experiments differing in the position of the excitation, the absolute values $|u_\omega|^2 = |v_\omega + iw_\omega|^2$ are shown in Fig. 5.4. For both the frequency is $\omega = 50$. While the waves injected in the first experiment travel through the domain largely unaffected, those of the second are deflected at an a priori unknown scatterer.

For the inversion, we consider 24 experiments with 8 equidistantly spaced source positions and frequencies $\omega_i \in \{30, 40, 50\}$. For these frequencies, the wavelengths are between 0.125 and 0.21. The inclusion to be identified is a circle of radius 0.15 with $a = 1.3$ embedded in a material with $a = 1$.

Fig. 5.5 shows the identified coefficient for two cases. In the left, the two materials are known, so that sharp bounds $1 \leq a \leq 1.3$ can be posed. Instead, if we do not know the materials, we only use a rough guess $0.5 \leq a \leq 5$ and obtain the coefficient displayed on the right of the figure. No regularization was used in both cases.

In Fig. 5.6, the performance of the weighted error estimate (5.12) as a mesh refinement criterion is compared to global refinement and the $\eta^{\nabla\nabla u}$ indicator (2.11), which performed best after the weighted estimator for the Laplace equa-

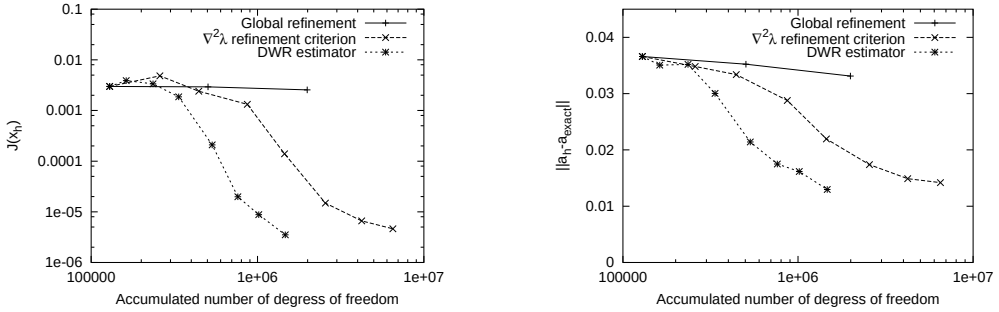


Figure 5.6: *Identification of an inclusion. Left: Reduction of target functional $J(x_h)$ for various refinement criteria, as function of the sum of the numbers of degrees of freedom of all 24 experiments. Right: Reduction of error $\|a - a_h\|$.*

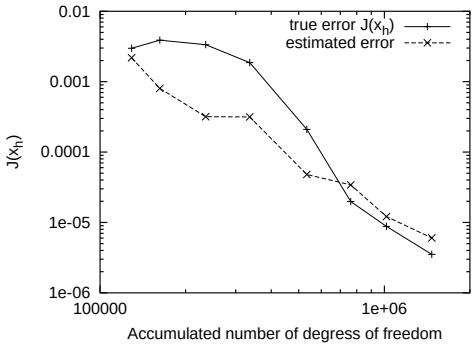


Figure 5.7: *Identification of an inclusion. Comparison of actual and estimated error.*

tion (see Section 2.1.3). As can be seen, the weighted indicator is significantly better than the other criteria, both in terms of reduction of the target functional $J(\cdot)$, and of the error $\|a_h - a_{exact}\|$ which is of greater practical interest. Thus, it is obvious that using this indicator can reduce the effort to solve the identification problem to a given accuracy greatly. Finally, Fig. 5.7 shows that the estimated errors in the target functional $J(\cdot)$ match the true ones reasonably good on finer meshes.

5.6 Application: Transmission tomography

As second example, we consider a similar layout as in the previous example, but for the much more challenging case that measurements are only available at the right boundary. This is the typical mode for so-called *cross-hole*, or *transmission tomography* in geophysics, where explosives are placed in one bore-hole, and receivers in a second hole a certain distance away.

A sketch of the layout of this example is given in Fig. 5.8. As an abstract description of this situation, we choose the same domain as in the previous

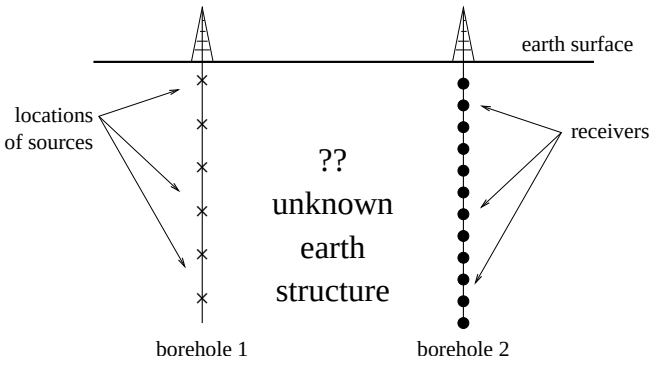


Figure 5.8: *Transmission tomography. Layout of example.*

example, but with homogeneous Neumann boundary conditions at the top to simulate the free boundary earth surface. At the left, the sources are represented by Dirichlet values, and at the right and bottom simple absorbing boundaries are given again, to indicate that these are artificial boundaries. The measurements are the Neumann values along the right borehole, i.e.

$$m(u_{\omega_i} - z) = \frac{1}{2}\|\partial_n u_{\omega_i} - z\|_\Gamma^2, \qquad \Gamma = \partial\Omega \cap \{x = 1\}.$$

The goal is the identification of the medium between the two boreholes. As an idealized situation, we choose the same coefficient structure as in the previous example, i.e. a circular inclusion, but with smaller variation $1 \leq a \leq 1.1$. The size of this variation in the coefficient is common for geophysical media. The setting of this example is comparable to that used by Pratt et al. [55], but we use a significantly higher resolution.

For the identification problem, we use 8 locations for sources along the left borehole, and the frequencies $\omega = \{20, 25, 30, 35\}$ at each location, making a total of 32 experiments.

As pointed out in Section 5.3, this problem is difficult since relatively small changes in the coefficient can shift the phase of the wave at the receiver positions by more than half a wavelength, leading to identification of a local minimum instead of the global one. Therefore, we start with the constant value 1.05, which is close enough for the identification process to find the global optimum. Nevertheless, this initial value does not reveal information about the structure of the sought coefficient. The problem is also challenging since it is necessary to solve the state equation to rather high accuracy.

The results of computations can be seen in Fig.s 5.9–5.11. In the first figure, the identified coefficient is shown. Its structure is clearly resolved, although the vertical extension of the inclusion is not computed accurately. However, given the limited amount of information used for the inversion, this resolution is already very good. Unfortunately, the computation could not be extended to higher numbers of degrees of freedom due to computational restrictions.

In the second figure, 5.10, the reduction of target functional $J(x_h)$ and the error $\|a_h - a_{exact}\|$ is shown for the same two refinement criteria as above, i.e. the

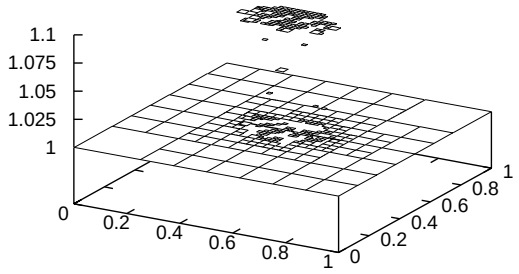


Figure 5.9: *Transmission tomography. Identified coefficient.*

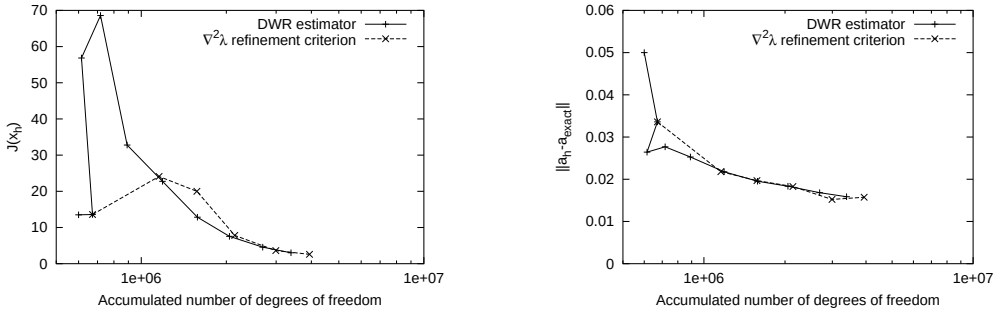


Figure 5.10: *Transmission tomography. Left: Reduction of target functional $J(x_h)$ for two refinement criteria, as function of the sum of the numbers of degrees of freedom of all 24 experiments. Right: Reduction of error $\|a - a_h\|$.*

weighted error estimate (5.12) and the $\eta^{\nabla\nabla u}$ indicator (2.11). Unlike in the previous application, but as for some of the cases discussed in Section 2.1.3, the weighted error indicator is not better than the one using second derivatives of the Lagrange multiplier. The weighted error estimator even shows an irregular behavior on coarse grids.

On the other hand, Figure 5.11 shows that estimated and true errors with respect to the target functional $J(\cdot)$ coincide almost perfectly and the ratio is very close to one, despite the initial irregular behavior of $J(x_h)$.

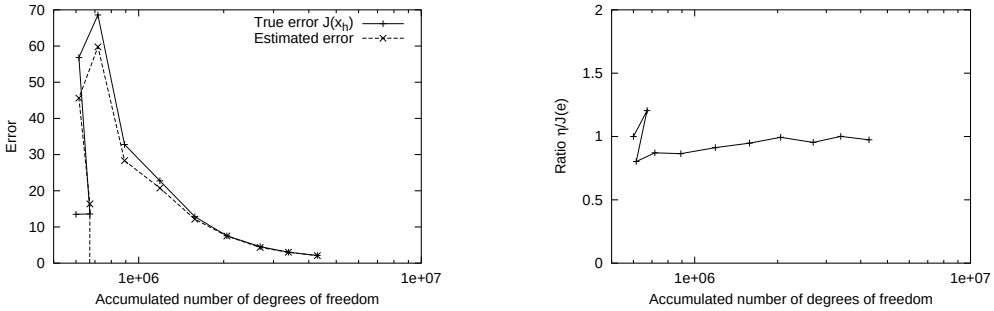


Figure 5.11: *Transmission tomography. Comparison of actual and estimated error. Left: Absolute values. Right: Ratio between estimated and true error.*

Outlook

Inverse problems associated with partial differential equations provide plenty of subjects for research. Efficient techniques for adaptivity, error estimation, and the treatment of bound constraints have been discussed in this thesis. However, at least the following three topics for further research immediately come to mind that have not or only briefly been touched in this work:

Uncertainty quantification and inversion for probability distributions. In this work, we have concentrated on solving for *one* parameter function that best describes the measured observation. However, this leaves one aspect entirely out of view: measurements are usually noisy, i.e. the measurement we have used is only one instance of a family of measurements that satisfy a certain probability law. From each possible noisy measurement there follows an inverted parameter to which we assign the same probability that the measurement has from which it was computed. The true solution of an inverse problem would therefore be a probability density in the space of parameters, i.e. a functional that assigns each element of the parameter space a probability value.

Knowing this probability distribution would give us enormous information. For example, it would be simple to assess the local or global resolution, i.e. the accuracy with which the parameter was resolved globally or at certain points of the domain. This would be necessary to evaluate the reliability of the solution. If we are not satisfied with the resolution, we could make more experiments. Knowledge of the probability distribution could also be used for *experimental design*, which tailors experimental set-ups such that they yield maximal resolution, again either globally or locally.

The downside, of course, is the likewise enormous complexity of the task. Little has been achieved in this field since the influential book by Tarantola [63] appeared in 1987 and made this aspect of inverse problems available to the greater public in applied sciences. In a few approaches (see, for example, Banks and Bihari [12] and Wojtkiewicz et al. [67]) the measurement space was sampled to obtain respective samples in parameter space, but by and large probability density recovery has been avoided for the practical solution of PDE constrained parameter estimation problems.

Truly inverting for the probability density beyond recovering half-widths in linear least squares problems with Gaussian noise offers an exciting field of research. With the recent advent of massively parallel clusters of workstations, the necessary computing power to solve the literally thousands or millions of forward problems seems already in place to do this for small problems.

Optimal choice of regularization. As O'Leary [52] puts it, "choosing the regularization parameter is an art based on good heuristics and prior knowledge of the noise in the observation". Although we have neglected this question entirely in this work, any reliable approach to inversion needs to have an automatic strategy for the selection of the regularization parameter. A large number of approaches for this exist, see for example the book by Engl, Hanke and Neubauer [32] on the subject. However, most of these approaches only have a theoretical foundation for linear problems and/or require the solution of a significant number of additional problems, and only few seem to be suited for the large scale nonlinear problems associated with partial differential equation.

One can probably say that these strategies have not yet found their way into the solution of nonlinear PDE constrained problems and the aspect of "art" and "heuristics" prevails to date. This calls for further research in the field. Duality and sensitivity as touched in this work could well be one building block for approaches for this. In particular they might help in an extension where we make the regularization parameter a space dependent function: set it to a large value where not enough information is available to recover the desired information, but set it to a small value where we have this information and do not need much regularization.

Efficient solution of large scale problems. Compared to some practical applications, the examples in this work are toy problems. Inversion of seismic data is frequently listed among the most computationally intensive applications presently solved in industry, for a good reason: it usually involves PDEs stated in three space and one time dimension, these PDEs are wave equations with high frequency solutions and are thus hard to solve, the number of measurements goes into the thousands, and the required resolution is high. Handling the amount of data, measurements in the range of many gigabytes, is a challenge in itself. The computational complexity of this task is not one or two orders of magnitude away from the examples in this work, but several.

Yet, the programs used in practice are algorithmically simple. They do not usually use adaptivity for the solution of the PDE, or include error estimation. They often do not even involve multiple experiment structures but invert for one dataset after the other. Combining the algorithms and mathematical methods of this work with practical applications is likely to gain a significant reduction of numerical effort, and an increase in resolving power. However, the expected complexity of such programs calls for a very careful design that in itself justifies research.

Outlook. It is probably safe to assert that PDE constrained inverse and optimization problems will become a major subject of research in the near future. Adaptive methods and error estimation will become as pervasive as they are now in the numerical solution of partial differential equation. Given the challenges and the potential practical applications, this promises to become an interesting field!

Bibliography

[1] Robert Acar. Identification of coefficients in elliptic equations. *SIAM J. Control Optim.*, 31:1221–1244, 1993.

[2] L. Amundsen. Comparison of the least-squares criterion and the Cauchy criterion in frequency-wavenumber inversion. *Geophysics*, 56:2027–2035, 1991.

[3] Uri M. Ascher and Eldad Haber. Grid refinement and scaling for distributed parameter estimation problems. *Inverse Problems*, 17:571–590, 2001.

[4] Uri M. Ascher and Eldad Haber. A multigrid method for distributed parameter estimation problems. Technical report, University of British Columbia, 2002.

[5] Wolfgang Bangerth. Adaptive Finite-Elemente-Methoden zur Lösung der Wellengleichung mit Anwendung in der Physik der Sonne. Thesis, University of Heidelberg, 1998.

[6] Wolfgang Bangerth. Mesh adaptivity and error control for a finite element approximation of the elastic wave equation. In Alfredo Bermúdez, Dolores Gómez, Christophe Hazard, Patrick Joly, and Jean E. Roberts, editors, *Proceedings of the Fifth International Conference on Mathematical and Numerical Aspects of Wave Propagation (Waves2000), Santiago de Compostela, Spain, 2000*, pages 725–729. SIAM, 2000.

[7] Wolfgang Bangerth. Multi-threading support in `deal.II`. Preprint 2000-11, SFB 359, Universität Heidelberg, April 2000.

[8] Wolfgang Bangerth. Using modern features of C++ for adaptive finite element methods: Dimension-independent programming in deal.II. In Michel Deville and Robert Owens, editors, *Proceedings of the 16th IMACS World Congress 2000, Lausanne, Switzerland, 2000*. IMACS – Department of Computer Science, Rutgers University, New Brunswick, 2000. Document Sessions/118-1.

[9] Wolfgang Bangerth, Ralf Hartmann, and Guido Kanschat. `deal.II` *Differential Equations Analysis Library, Technical Reference*. IWR, Universität Heidelberg, April 2002. `http://gaia.iwr.uni-heidelberg.de/~deal/`.

[10] Wolfgang Bangerth and Guido Kanschat. Concepts for object-oriented finite element software – the `deal.II` library. Preprint 99-43, SFB 359, Universität Heidelberg, October 1999.

[11] Wolfgang Bangerth and Rolf Rannacher. Adaptive finite element techniques for the acoustic wave equation. *J. Comput. Acoustics*, 9(2):575–591, 2001.

[12] H. T. Banks and Kathleen L. Bihari. Modelling and estimating uncertainty in parameter estimation. *Inverse Problems*, 17:95–111, 2001.

[13] H. T. Banks and K. Kunisch. *Estimation Techniques for Distributed Parameter Systems*. Birkhäuser, Basel–Boston–Berlin, 1989.

[14] A. Bayliss and E. Turkel. Radiation boundary conditions for wave-like equations. *Commun. Pure Appl. Math.*, 33:707–725, 1980.

[15] R. Becker. Adaptive finite elements for optimal control problems. Habilitation thesis, University of Heidelberg, 2001.

[16] R. Becker, H. Kapp, and R. Rannacher. Adaptive finite element methods for optimal control of partial differential equations: Basic concept. *SIAM J. Contr. Optim.*, 39:113–132, 2000.

[17] R. Becker and R. Rannacher. An optimal control approach to error estimation and mesh adaptation in finite element methods. *Acta Numerica*, 10:1–102, 2001.

[18] Roland Becker and Boris Vexler. Mesh adaptation for parameter identification problems. Proceedings of ENUMATH 2001, 2002. submitted.

[19] Hend Ben Ameur, Guy Chavent, and Jérôme Jaffre. Raffinement et déraffinement de paramétrisation pour l'estimation de transmissivité hydraulique. Rapport de recherche 3623, INRIA, 1999.

[20] Maïtine Bergounioux, Kazufumi Ito, and Karl Kunisch. Primal-dual strategy for constrained optimal control problems. *SIAM J. Control Optim.*, 37:1176–1194, 1999.

[21] H. Blum and F.-T. Suttmeier. Weighted error estimates for finite element solutions of variational inequalities. *Computing*, 65:119–134, 2000.

[22] H. G. Bock. Numerical treatment of inverse problems in chemical reaction kinetics. In K. H. Ebert, P. Deuflhard, and W. Jäger, editors, *Modelling of Chemical Reaction Systems*, volume 18 of *Springer Series in Chemical Physics*. Springer, Heidelberg, 1981.

[23] H. G. Bock. *Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen*, volume 183 of *Bonner Mathematische Schriften*. University of Bonn, Bonn, 1987.

[24] Guy Chavent. Duality methods for waveform inversion. Rapport de recherche 2975, INRIA, 1996.

[25] Guy Chavent and Robert Bissell. Indicators for the refinement of parameterization. In M. Tanaka and G.S. Dulikravich, editors, *Inverse Problems in Engineering Mechanics (Proceedings of the third International Symposium on Inverse Problems ISIP'98 held in Nagano, Japan)*, pages 309–314. Elsevier, 1998.

[26] Guy Chavent, Karl Kunisch, and Jean E. Roberts. Primal-dual formulations for parameter estimation problems. *Comp. Appl. Math.*, 18:173–229, 1999.

[27] Zhiming Chen and Jun Zou. An augmented Lagrangian method for identifying discontinuous parameters in elliptic systems. *SIAM J. Control Optim.*, 37:892–910, 1999.

[28] P. Deuflhard, H. W. Engl, and O. Scherzer. A convergence analysis of iterative methods for the solution of nonlinear ill-posed problems under affinely invariant conditions. *Inverse Problems*, 14:1081–1106, 1998.

[29] M. Diehl, H.G. Bock, and J.P. Schlöder. Newton type methods for the approximate solution of nonlinear programming problems in real-time. Technical report, University of Heidelberg, 2001. to appear.

[30] I. S. Duff and J. K. Reid. The multifrontal solution of indefinite sparse symmetric linear equations. *ACM Trans. Math. Softw.*, 9:302–325, 1983.

[31] I. S. Duff and J. K. Reid. MA47, A Fortran code for the direct solution of indefinite sparse symmetric systems. Technical Report RAL-95-001, Rutherford Appleton Laboratory, Oxfordshire, UK, 1995.

[32] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, Dordrecht, 1996.

[33] Bjorn Engquist and Andrew Majda. Absorbing boundary conditions for the numerical simulation of waves. *Math. Comp.*, 31:629–651, 1977.

[34] Richard S. Falk. Error estimates for the numerical identification of a variable coefficient. *Math. Comp.*, 162:537–546, 1983.

[35] Colin G. Farquharson and Douglas W. Oldenburg. Non-linear inversion using general measures of data misfit and model structure. *Geophys. J. Int.*, 134:213–227, 1998.

[36] Roland W. Freund and Noel M. Nachtigal. A new Krylov-subspace method for symmetric indefinite linear systems. In W. F. Ames, editor, *Proceedings of the 14th IMACS World Congress on Computational and Applied Mathematics*, pages 1253–1256, 1994.

[37] E. Haber and U. M. Ascher. Preconditioned all-at-once methods for large, sparse parameter estimation problems. *Inverse Problems*, 17:1847–1864, 2001.

[38] E. Haber, U. M. Ascher, and D. Oldenburg. On optimization techniques for solving nonlinear inverse problems. *Inverse Problems*, 16:1263–1280, 2000.

[39] Eldad Haber and Douglas Oldenburg. Joint inversion: a structural approach. *Inverse Problems*, 13:63–77, 1997.

[40] The Harwell Subroutine Library. http://www.cse.clrc.ac.uk/Activity/HSL.

[41] K. Ito, M. Kroller, and K. Kunisch. A numerical study of an augmented Lagrangian method for the estimation of parameters in elliptic systems. *SIAM J. Sci. Stat. Comput.*, 12:884–910, 1991.

[42] Kazufumi Ito and Karl Kunisch. The augmented Lagrangian method for parameter estimation in elliptic systems. *SIAM J. Contr. Optim.*, 28:113–136, 1990.

[43] Kazufumi Ito and Karl Kunisch. Sensitivity analysis of solutions to optimization problems in Hilbert spaces with applications to optimal control and estimation. *J. Diff. Eq.*, 99:1–40, 1992.

[44] Kazufumi Ito and Karl Kunisch. Augmented Lagrangian-SQP-methods in Hilbert spaces and application to control in the coefficients problems. *SIAM J. Optimization*, 6:96–125, 1996.

[45] Kazufumi Ito and Karl Kunisch. Estimation of the convection coefficient in elliptic equations. *Inverse Problems*, 13:995–1013, 1997.

[46] Barbara Kaltenbacher. A projection-regularized Newton method for nonlinear ill-posed problems and its application to parameter identification problems with finite element discretization. *SIAM J. Numer. Anal.*, 37:1885–1908, 2000.

[47] C. Kravaris and J. H. Seinfeld. Identification of parameters in distributed parameter systems by regularization. *SIAM J. Control Optim.*, 23:217–241, 1985.

[48] Matti Lassas, Margaret Cheney, and Gunther Uhlmann. Uniqueness for a wave propagation inverse problem in a half space. *Inverse Problems*, 14:679–684, 1998.

[49] R. Luce and S. Perez. Parameter identification for an elliptic partial differential equation with distributed noisy data. *Inverse Problems*, 15:291–307, 1999.

[50] H. Maurer and J. Zowe. First and second order necessary and sufficient conditions for infinite-dimensional programming problems. *Math. Programming*, 16:98–110, 1979.

[51] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization.* Springer Series in Operations Research. Springer, New York, 1999.

[52] Dianne O'Leary. Near-optimal parameters for Tikhonov and other regularization schemes. *SIAM J. Sci. Comput.*, 23:1161–1171, 2001.

[53] C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.*, 12:617–629, 1975.

[54] R. L. Parker. *Geophysical Inverse Theory.* Princeton University Press, Princeton, NJ, 1994.

[55] R. Gerhard Pratt, Changsoo Shin, and G. J. Hicks. Gauss-Newton and full Newton methods in frequency-space seismic waveform inversion. *Geophys. J. Int.*, 133:341–362, 1998.

[56] Yousef Saad. *Iterative Methods for Sparse Linear Systems.* PWS, 1996.

[57] J. P. Schlöder. *Numerische Methoden zur Behandlung hochdimensionaler Aufgaben der Parameteridentifizierung*, volume 187 of *Bonner Mathematische Schriften.* University of Bonn, Bonn, 1988.

[58] Volker H. Schulz. Solving discretized optimization problems by partially reduced SQP methods. *Comput. Visual. Sci.*, 1:83–96, 1998.

[59] C. Schwab. *p- and hp-Finite Element methods. Theory and Applications to Solid and Fluid Mechanics.* Oxford University Press, 1998.

[60] M. K. Sen and P. L. Stoffa. Nonlinear one-dimensional seismic waveform inversion using simulated annealing. *Geophysics*, 56:1624–1638, 1991.

[61] Franz-Theo Suttmeier. *Error Analysis for Finite Element Solutions of Variational Inequalities. Professorial Dissertation.* University of Dortmund, 2001.

[62] Franz-Theo Suttmeier. General approach for a posteriori error estimates for finite element solutions of variational inequalities. *Comp. Mech.*, 27:317–323, 2001.

[63] Albert Tarantola. *Inverse Problem Theory.* Elsevier, 1987.

[64] Ulrich Tautenhahn. Error estimates for regularized solutions of non-linear ill-posed problems. *Inverse Problems*, 10:485–500, 1994.

[65] Stefan Turek. *Efficient Solvers for Incompressible Flow Problems.* Number 6 in Lecture Notes in Computational Science and Engineering. Springer, 1999.

[66] S. P. Vanka. Implicit multigrid solutions of Navier–Stokes equations in primitive variables. *J. Comput. Phys.*, 65:138–158, 1985.

[67] S. F. Wojtkiewicz, M. S. Eldred, R. V. Field, A. Urbina, and J. R. RedHorse. Uncertainty quantification in large computational engineering models. Technical Report AIAA-2001-1455, Sandia National Laboratories, 2001.

[68] Kôsaku Yosida. *Functional Analysis.* Springer, 1980.

[69] W. H. Yu. Necessary conditions for optimality in the identification of elliptic systems with parameter constraints. *J. Optimization Theory Appl.*, 88:725–742, 1996.