JYVÄSKYLÄ STUDIES IN COMPUTING
132

Olli Mali

Analysis of Errors Caused by Incomplete Knowledge of Material Data in Mathematical Models of Elastic Media

JYVÄSKYLÄN YLIOPISTO

JYVÄSKYLÄ STUDIES IN COMPUTING 132

Olli Mali

Analysis of Errors Caused by Incomplete Knowledge of Material Data in Mathematical Models of Elastic Media

Esitetään Jyväskylän yliopiston informaatioteknologian tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Ylistönrinteellä, salissa YAA303
syyskuun 16. päivänä 2011 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Information Technology of the University of Jyväskylä,
in Ylistönrinne, hall YAA303, on September 16, 2011 at 12 o'clock noon.

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2011

Analysis of Errors Caused by Incomplete Knowledge of Material Data in Mathematical Models of Elastic Media

JYVÄSKYLÄ STUDIES IN COMPUTING 132

Olli Mali

Analysis of Errors Caused by Incomplete Knowledge of Material Data in Mathematical Models of Elastic Media

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2011

Editors
Timo Männikkö
Department of Mathematical Information Technology, University of Jyväskylä
Pekka Olsbo, Ville Korkiakangas
Publishing Unit, University Library of Jyväskylä

URN:ISBN:978-951-39-4394-3
ISBN 978-951-39-4394-3 (PDF)

ISBN 978-951-39-4393-6 (nid.)
ISSN 1456-5390

Jyväskylä University Printing House, Jyväskylä 2011

ABSTRACT

Mali, Olli
Analysis of errors caused by incomplete knowledge of material data in mathematical models of elastic media
Jyväskylä: University of Jyväskylä, 2011, 111 p.
(Jyväskylä Studies in Computing
ISSN 1456-5390; 132)
ISBN 978-951-39-4393-6 (nid.), 978-951-39-4394-3 (PDF)
Finnish summary
Diss.

We study the effects that incompletely known data introduce to problems in continuum mechanics. In particular, we are interested in the case when the parameters of the media in the constitutive laws are not completely known. Our analysis is based on deviation estimates, which are functionals that allow us to study the distance between an arbitrary function from the energy space and the exact solution of the problem. For the Kirchhoff–Love arch model, deviation estimates are derived for the first time. Since the data are not unique, we have a set of solutions instead of a single function. For a certain class of problems, we present estimates for the radius of the solution set, where estimates depend on the problem data and the accuracy by which the data are known. For a linear isotropic elasticity problem, we show that for a certain type of boundary conditions the elastic energy becomes very sensitive to small variations in Poisson's ratio at the incompressibility limit. This phenomenon may be significant enough to render quantitative analysis meaningless even relatively far from the limit.

Keywords: indeterminate data, functional deviation estimates, a posteriori error estimates, partial differential equations, linear elasticity

Author Olli Mali
Department of Mathematical Information Technology
University of Jyväskylä
Finland

Supervisors Professor Pekka Neittaanmäki
Department of Mathematical Information Technology
University of Jyväskylä
Finland

Professor Sergey Repin
Steklov Institute, St. Petersburg department
Russian Academy of Sciences
Russia
Department of Mathematical Information Technology
University of Jyväskylä
Finland

Reviewers Professor Ulrich Langer
Institute of Computational Mathematics
Johannes Kepler University Linz
Austria

Dr. Dan Tiba
Institute of Mathematics
Romanian Academy
Romania

Opponent Professor Rolf Stenberg
Institute of Mathematics
Aalto University
Finland

ACKNOWLEDGEMENTS

I started to work under the supervision of Prof. Pekka Neittaanmäki and Prof. Sergey Repin in early 2007. The research topic felt immediately interesting and well motivated. During these years, they supported my scientific work and much more. They have introduced the modern scientific world to me and set an example of a true scientist driven by the passion for understanding the world we live in.

I am very grateful for my reviewers Prof. Ulrich Langer and Dr. Dan Tiba for their valuable comments and the interest they showed towards the Thesis. In particular, I want to thank Prof. Langer for his kind invitation to visit Radon Institute and Dr. Tiba for the great effort he put to improve the quality of the Thesis during his visit in Jyväskylä.

I want to acknowledge that this work was partially supported by the FICS graduate school.

I also want to express my gratitude for all my colleagues working at the University of Jyväskylä for their good company. In particular, I want to mention Juha Jeronen, whose computer skills have been invaluable more than once. Also I want to thank Immanuel Anjam from our research team for our discussions. I want to thank Tero Tuovinen, Marja–Leena Rantalainen and Kati Valpe as well as all the secretaries of the department for helping me with countless practical issues.

I also want to thank Euler Institute for their invitations to St. Petersburg. With the precious help from local members of our research team, namely Svetlana Matsulevits and Elena Gorshkova my visits there have been a pleasant experience.

My scientific career leading to the Thesis was started in a certain basement on spring 1999. I want to thank Juhana Ahtiainen for his support then and now, as he helped to proofread the Thesis. I also want to thank all my other good friends, who have made the time of study enjoyable.

I want to thank my parents and family, who have supported me on everything I have done. Last and most importantly, I want to thank my lovely wife, Julia. She may not have solely accelerated the production of the Thesis, but has made that time full of joy and love.

Olli Mali
Jyväskylä, August 2011

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

If a mathematical model is used to predict the outcome of some scenario, then it is important to understand that the incomplete knowledge of the data affects the accuracy of any results the model produces. We should distinguish between natural laws, which we assume to hold exactly in the model, such as Newton's laws, conservation of momentum, etc. and laws that are constructed by experimental means or are based on simplifying assumptions. Examples of the latter, and the type mainly studied in the Thesis, are various constitutive relations in media. Moreover, the initial uncertainty may cumulate in an unexpected way. This is of practical importance in engineering applications that aim to construct designs based on results obtained from mathematical models.

The aim of the Thesis is to investigate, with the paradigm of certain models generated by partial differential equations (PDE), how this uncertainty may affect quantitative analysis. This knowledge is important for the following two reasons:

- Typically, a mathematical model can not be solved exactly, but approximate solutions are constructed. The difference between the approximate solution and the exact solution is called *approximation error*. Reduction of this error typically requires additional computational effort and resources. Due to the presence of the inaccuracy arising from incompletely known natural laws, there is an *accuracy limit* beyond which additional computations make no sense.

- In engineering practice, results obtained from mathematical models often serve as a basis for decision making. Unawareness of uncertainty in simulation results creates a dangerous false sense of reliability.

The main approach usually used to control uncertainty in a model is the so called probabilistic approach, which leads to stochastic PDEs. There indeterminate data entering a PDE are considered random variables with a known probability density. The aim is then to find or approximate the mean value, variance and other probabilistic quantities related to the solution. An overview of the theory and related numerical methods (dating back to [3]) can be found in [46]. The most popular numerical method is the Monte Carlo method [10]. The idea of the method is to generate samples of input data within an uncertainty range and to compute respective solutions. Then the probabilistic features are studied via a statistical analysis.

Probability distributions are not the only way to model uncertainty. In the evidence theory (also known as Dempster-Shafer theory) [7, 47], the requirements for the probability measure are relaxed and obtained probability assignments are applied instead.

In the theory of fuzzy sets [56] (and evolved possibility theory [57]), the uncertainty is introduced via a membership function. In the classical theory, there are only two options, an element is a member of a set or it is not. The membership function defines the degree of truth of the statement that the element belongs to a set. The idea is to analyze how the fuzziness of the data is inherited in to the solution. An introduction and examples can be found in [5].

The application of these theories to physical models is discussed in [19], where uncertainty is studied in the framework of the worst–case scenario method.

A branch of mathematics called reliability engineering is related to system analysis and risk analysis. An overview is given in [59], where uncertainty is classified as either aleatory or epistemic. The first refers to the uncertainty due to inherent variability of the system and the latter to a lack of knowledge.

Another concept related to incompletely known data is the sensitivity analysis. It indicates how severe is the influence of the perturbation of a particular input parameter on the solution or other quantity of interest. Typically, the derivative of the solution with respect to the input parameters is investigated. The analysis can be done for the original PDE or the approximated finite dimensional model. For an exposition of the corresponding theory see [16, 23, 45]. The sensitivity analysis can also be done numerically by Monte Carlo type simulations, where the scattering of the results may indicate the level of correlation between input and output data (see, e.g., [20]). The sensitivity of the solution with respect to the geometry of the problem is studied in [17, 18], where the information of sensitivity is used to solve the optimization problem.

In the Thesis, a certain kinds of elliptic boundary value problems are analyzed. The physical and mathematical background required for the indeterminacy analysis is presented in Chapter 2. The physical models are presented in Section 2.1 and their mathematical generalization in Section 2.2. In Section 2.2, the existence and uniqueness of the solutions for problems of this type are discussed.

Our analysis is based on functional estimates, which estimate the distance between the exact solution of a PDE and an arbitrary admissible function from the energy space for a wide class of problems, as discussed in two books [32, 36]. These estimates (unlike other a posteriori estimates for PDEs) are derived purely on a functional level and do not contain any mesh–dependent constants. They are known as a posteriori error estimates of the functional type or deviation estimates. This methodology is applicable not only to control the approximation error, but also to estimate the modelling error [41, 42] and the error caused by an incomplete knowledge of data [32, 38]. The deviation estimates applied in the

Thesis are presented and derived in Section 2.3. In Chapter 3, the derivation is presented in the paradigm of the Kirchhoff–Love arch problem.

Our main emphasis is to demonstrate how the deviation estimates can be used to analyze the effects of uncertain data. The indeterminacy analysis is based on the following key properties of the deviation estimates:

(a) They depend *explicitly* on the problem data,

(b) They are guaranteed estimates in the sense that upper bound is always greater or equal, and lower bound is lower or equal, than the true deviation,

(c) They are computable and independent of any approximation method. Instead, they are derived on a functional level and can be interpreted as an alternative formulation of the problem.

These properties lead to the analysis done in Chapter 4, where the effects of uncertain data are discussed. In this Chapter the problem statement is as follows: Let the problem data belong to a set, instead of being exactly known; then the problem possesses a set of solutions instead of a single solution. The relations between the set of data and the set of solutions are of practical interest. In Section 4.1 we demonstrate some effects of indeterminate data and define various relations with the help of simple algebraic examples. The relation between the indeterminate data and the solution set was studied for diffusion type problems with uncertain coefficients in [26, 27, 28].

From the engineering point of view, it is very important to know how limited accuracy of material parameters affects the values of stress and displacement or other critical design parameters. The special case of isotropic linear elasticity is discussed in Section 4.4. We derive asymptotic estimates, which demonstrate that the solution energy can be very sensitive to small changes in Poisson's ratio close to the incompressibility limit, depending on whether the boundary conditions belong to a certain class or not. Moreover, by a problem that admits an analytical solution we demonstrate that the inaccuracy generated by an incompletely known Poisson's ratio may render a quantitative analysis meaningless relatively far from the incompressibility limit.

The results presented in the Thesis can be listed as follows:

(a) The two–sided estimates for the radius of the solution set for the elliptic problem with uncertain coefficients are presented. The main result is Theorem 4.1, which is directly applicable to the majority of physical models presented in Section 2.1. Moreover, we demonstrate that the derived upper bound for the radius of solution set is exact. This fact is demonstrated with the paradigm of a model problem, where the upper bound is indeed achieved.

(b) The deviation estimates for the Kirchhoff–Love arch problem are derived.

(c) The conditions for the blow–up of the indeterminacy caused by an incompletely known Poisson's ratio for the isotropic linear elasticity problem are presented. We identify the asymptotic behavior of the logarithmic derivative of the energy with respect to small variations of Poisson's ratio. The connection between the asymptotic behavior and the space of admissible solutions (boundary conditions) is explicitly given.

The implications that these results have to the simulations and computational practice is discussed in Chapter 5.

Author's contribution

The topics exposed in the Thesis have been suggested by the two supervisors and are based on their earlier work related to deviation estimates. The deviation estimates for Kirchhoff–Love arches presented in Chapter 3 (published in [24]) were derived independently by the author with only advisory help.

The results on estimates of the solution set for linear models and results on blow up for the isotropic linear elasticity model exposed in Chapter 4 were mainly obtained by intensive joint work of the author and Prof. S. Repin, during which we discussed results and ideas with Prof. P. Neittaanmäki. The results of Chapter 4 have been partly published in [25, 26, 27, 28, 29]. However, the published results concerning the radius of the solution set were obtained by the paradigm of the diffusion problem, whereas in the Thesis they are derived for a generalized variational problem and the properties of the estimates are discussed in more detail.

The author was also a member of a group of scientists (together with I. Anjam, A. Muzalevski, and the supervisors), who proposed and implemented certain types of a posteriori error estimates for a Maxwell type problem in [2].

2 MATHEMATICAL AND PHYSICAL BACKGROUND

The space of d–dimensional real valued vectors is denoted by $\mathbb{R}^d$ and $\mathbb{M}^{d\times d}$ denotes the space of $d \times d$ matrices (second order tensors). Their scalar (inner) products in $\mathbb{R}^d$ and $\mathbb{M}^{d\times d}$ are defined as

$$x \cdot z = \sum_{i=1}^{d} x_i z_i \quad \text{and} \quad \sigma : \tau = \sum_{i=1}^{d} \sum_{j=1}^{d} \sigma_{ij} \tau_{ij}$$

respectively. They generate the norms

$$|x| = \sqrt{x \cdot x} \quad \text{and} \quad |\sigma| = \sqrt{\sigma : \sigma}.$$

We recall that Banach space is a (vector) space where all Cauchy–sequences converge to an element of the space. Hilbert space U is a Banach space where norm is defined by an inner product. Let $f, g \in U$, the inner product is a mapping $U \times U \to \mathbb{R}$. It is denoted by parenthesis (f, g). Important Hilbert space is the space of real(vector/tensor) –valued square integrable functions (in the Lebesque sense) $L_2(\Omega)$, defined by the inner product

$$(f, g) := \int_{\Omega} f \cdot g \, \mathrm{dx}$$

and the respective norm

$$\|g\| = \sqrt{\int_{\Omega} |g|^2 \, \mathrm{dx}}.$$

We assume that Ω in $\mathbb{R}^d$ is a bounded connected domain with a Lipschitz continuous boundary Γ. The partial derivative of the function f is denoted by $\frac{\partial f}{\partial x_i}$ and defined in the weak sense through the integral relation

$$\int_{\Omega} \frac{\partial f}{\partial x_i} \cdot w \, \mathrm{dx} = -\int_{\Omega} f \cdot \frac{\partial w}{\partial x_i} \, \mathrm{dx}, \quad \forall w \in C_0^1(\Omega),$$

where $C_0^1(\Omega)$ is the space of once differentiable functions in Ω that vanish on the boundary Γ. The generalized (Sobolev, weak) derivative of order $|\alpha| : \sum \alpha_i$ is

$$D^\alpha v := \frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}},$$

where $\alpha = \alpha_1, \alpha_2, \dots, \alpha_d$ is the multi index.

Important differential operators applied in the Thesis are listed below:
The gradient of a scalar valued function w is a vector

$$\nabla w := \left[\frac{\partial w}{\partial x_i} \right]_{i=1}^d \in \mathbb{R}^d.$$

Similarly we denote the Jacobian of a vector valued function $[g_i]_{i=1}^d$

$$\nabla g := \left[\frac{\partial g_i}{\partial x_j} \right]_{i,j=1}^d \in \mathbb{M}^{d\times d}.$$

For a vector– and tensor–valued function the divergence is defined by

$$\operatorname{div} g := \sum_{i=1}^d \frac{\partial g_i}{\partial x_i}$$

and

$$\operatorname{Div} \tau := \left[\sum_{i=1}^d \frac{\partial \tau_{ij}}{\partial x_i} \right]_{j=1}^d \in \mathbb{R}^d,$$

respectively. In two- and three-dimensional spaces, the operator curl represents the rotation of the vector field. In $d = 3$, we have

$$\operatorname{curl} g := \left[\frac{\partial g_3}{\partial x_2} - \frac{\partial g_2}{\partial x_3}, \frac{\partial g_1}{\partial x_3} - \frac{\partial g_3}{\partial x_1}, \frac{\partial g_2}{\partial x_1} - \frac{\partial g_1}{\partial x_2} \right] \in \mathbb{R}^3.$$

In $d = 2$, we define two curl operators: for a vector function g it is

$$\operatorname{curl} g := \frac{\partial g_1}{\partial x_2} - \frac{\partial g_2}{\partial x_1} \in \mathbb{R}$$

and for a scalar function w it is

$$\overline{\operatorname{curl}} w := \left[\frac{\partial w}{\partial x_2}, -\frac{\partial w}{\partial x_1} \right] \in \mathbb{R}^2.$$

In the theory of linear elasticity, the tensor of a small strain is defined as the symmetric part of the tensor ∇w, i.e.,

$$\varepsilon(w) = \tfrac{1}{2} \left(\nabla w + (\nabla w)^T \right).$$

Henceforth, we use the following well known identities:

$$\operatorname{div}(wg) = g \cdot \nabla w + w\operatorname{div} g \tag{2.1}$$

and similarly for a vector valued function and a tensor,

$$\operatorname{Div}(\tau g) = \tau^T : \nabla g + g \cdot \operatorname{Div}\tau. \tag{2.2}$$

We recall the Ostrogradski integration formula

$$\int_\Omega \operatorname{div} g \, \mathrm{dx} = \int_\Gamma g \cdot n \, \mathrm{ds}. \tag{2.3}$$

From (2.1) and (2.3),

$$\int_\Omega \operatorname{div}(wg) \, \mathrm{dx} = \int_\Omega (g \cdot \nabla w + w\operatorname{div} g) \, \mathrm{dx} = \int_\Gamma wg \cdot n \, \mathrm{ds}, \tag{2.4}$$

where Γ is a closed surface of Ω and n is the outer normal of Γ.

Sobolev spaces

We use a standard definition for the Sobolev spaces $H^m(\Omega)$, which contains L_2-functions having generalized derivatives up to the order m ($m \geq 1$) in L_2, i.e.,

$$H^m(\Omega) := \{v \in L_2(\Omega) \mid D^\alpha v \in L_2(\Omega), \quad \forall m : |\alpha| \leq m\}.$$

The norm is defined by

$$\|f\|_{m,2,\Omega} := \left(\int_\Omega \sum_{|\alpha| \leq m} |D^\alpha f|^2 \, \mathrm{dx} \right)^{\frac{1}{2}}.$$

These spaces have subspaces denoted by subscript 0, in which the functions vanish on the part of the boundary where the Dirichlet condition is defined.

Similarly, spaces associated to differential operators are defined by the relations

$$\begin{aligned}
H(\Omega, \operatorname{div}) &:= \{v \in L_2(\Omega, \mathbb{R}^d) \mid \operatorname{div} v \in L_2(\Omega)\}, \\
H(\Omega, \operatorname{Div}) &:= \{v \in L_2(\Omega, \mathbb{M}^{d \times d}) \mid \operatorname{Div} v \in L_2(\Omega, \mathbb{R}^d)\}, \\
H(\Omega, \operatorname{curl}) &:= \{v \in L_2(\Omega, \mathbb{R}^d) \mid \operatorname{curl} v \in L_2(\Omega)\}
\end{aligned}$$

with the respective inner products

$$\begin{aligned}
(f, g)_{\mathrm{div}} &= \int_\Omega (f \cdot g + \operatorname{div} f \operatorname{div} g) \, \mathrm{dx}, \\
(\tau, \sigma)_{\mathrm{Div}} &= \int_\Omega (\tau : \sigma + \operatorname{Div}\tau \cdot \operatorname{Div}\sigma) \, \mathrm{dx}, \\
(f, g)_{\mathrm{curl}} &= \int_\Omega (f \cdot g + \operatorname{curl} f \cdot \operatorname{curl} g) \, \mathrm{dx},
\end{aligned}$$

and norms $\|\cdot\|_{\text{div}}$, $\|\cdot\|_{\text{Div}}$, and $\|\cdot\|_{\text{curl}}$.

Functions in Sobolev spaces have counterparts on Γ that form spaces of *traces*. For example, the operator $\gamma : H^1(\Omega) \to L_2(\Gamma)$ is called the trace operator if it satisfies the following conditions:

$$\begin{aligned} \gamma w &= w|_\Gamma, \quad \forall w \in C^1(\Omega), \\ \|\gamma w\|_{2,\Gamma} &\le c_{\Gamma\gamma}\|w\|_{1,2,\Omega}, \end{aligned}$$

where $c_{\Gamma\gamma}$ is a positive constant independent of w. The trace operator is a natural generalization of the trace defined for a continuous operator (in the pointwise sense). The image of γ in $L_2(\Gamma)$ is denoted by $H^{1/2}(\Gamma)$. Thus, $\gamma \in \mathcal{L}(H^1(\Omega), H^{1/2}(\Gamma))$.

Using the operator γ, we can define subspaces of Sobolev space V generated by functions vanishing on Γ or on some part $\Gamma_1 \subset \Gamma$ of positive $d-1$ measure,

$$V_0 := \{w \in V \mid \gamma w = 0, \text{ a.e. on } \Gamma_1\}.$$

Henceforth, we understand the boundary values in the sense of traces. The phrase "$u = \phi$ on Γ" means that the trace γu of a function u defined in Ω coincides with the given function ψ defined on Γ. If u and v are defined in Ω the phrase "$u = v$ on Γ" means that $\gamma(u - v) = 0$ on Γ. For the sake of simplicity, we later on omit γ.

Inequalities

Next, we recall some known inequalities (see, e.g., [22, 48]), which we use in the subsequent analysis. The Hölder inequality has the integral form

$$\int\limits_\Omega fg \,\mathrm{dx} \le \left(\int\limits_\Omega |f|^p \,\mathrm{dx}\right)^{\frac{1}{p}} \left(\int\limits_\Omega |g|^q \,\mathrm{dx}\right)^{\frac{1}{q}} \tag{2.5}$$

and the discrete one:

$$|a \cdot b| \le \left(\sum_{i=1}^d |a_i|^p \,\mathrm{dx}\right)^{\frac{1}{p}} \left(\sum_{i=1}^d |b_i|^q \,\mathrm{dx}\right)^{\frac{1}{q}}, \tag{2.6}$$

where $\frac{1}{p} + \frac{1}{q} = 1$ and $p, q > 0$. In the case $p = q = 2$, the inequality is the Cauchy–Schwartz inequality for the space $L_2(\Omega)$. The triangular inequality for functions is the Minkowski inequality

$$\|f + g\| \le \|f\| + \|g\|.$$

Similar inequalities hold for the vector and tensor spaces.

It is easy to show that for any function $f : X \times Y \to \mathbb{R}$

$$\sup_{x\in X} \inf_{y\in Y} f(x, y) \le \inf_{y\in Y} \sup_{x\in X} f(x, y). \tag{2.7}$$

is valid.

For the functions in $H_0^1(\Omega)$, the Friedrichs inequality

$$\|w\| \leq c_F \|\nabla w\|, \quad \forall w \in H_0^1(\Omega) \tag{2.8}$$

is valid. It belongs to an important class of inequalities. The importance of these inequalities will become clear in Section 2.2, where the existence of solutions for variational problems is discussed. A systematic analysis of constants associated with embedding inequalities is presented in [30].

Within the theory of linear elasticity, the analog of the Poincare-Friedrich type inequality is Korn's inequality (see, e.g., [8]) that establishes the equivalence of the norms

$$\|w\|^2_{1,2,\Omega} := \int_\Omega (|\nabla w|^2 + |w|^2) dx$$

and

$$[|w|]^2_{1,2,\Omega} := \int_\Omega (|\varepsilon(w)|^2 + |w|^2) dx.$$

From Korn's inequality follows the existence of a positive constant c_K such that

$$\|\nabla w\| \;\leq\; c_K \|\varepsilon(w)\| \qquad \forall w \in V_0, \tag{2.9}$$

where c_K is independent of w.

A practical algebraic inequality is Young's inequality,

$$|a+b|^2 \leq (1+\beta)|a|^2 + \left(1 + \frac{1}{\beta}\right) |b|^2, \quad \forall \beta > 0, \tag{2.10}$$

which is valid for a, b being scalars, vectors, or tensors.

2.1 Physical models

The problems presented here are motivated by the energy principles observed in nature. They all have quadratic energy functionals, which possess unique minimizers. In Section 2.2, we discuss these models in the framework of a unified approach.

2.1.1 The stationary heat equation

We consider a stationary heat equation with homogeneous Dirichlet boundary conditions,

$$\begin{aligned} -\operatorname{div} A \nabla u &= f, \quad \text{in } \Omega, \\ u &= 0, \quad \text{on } \Gamma. \end{aligned}$$

The problem has a variational form generated by the energy functional

$$J(w) = \frac{1}{2} \int_{\Omega} A\nabla w \cdot \nabla w \, \mathrm{dx} - \int_{\Omega} f w \, \mathrm{dx}, \tag{2.11}$$

where $A \in L_\infty(\Omega, \mathbb{M}^{d\times d})$ is symmetric and positive definite. The generalized solution $u \in H_0^1(\Omega)$ satisfies the integral relation

$$\int_{\Omega} A\nabla u \cdot \nabla w \, \mathrm{dx} = \int_{\Omega} f w \, \mathrm{dx}, \quad \forall w \in H_0^1(\Omega). \tag{2.12}$$

The problem can be decomposed to two physical relations,

$$-\mathrm{div} p = f, \tag{2.13}$$
$$p = A\nabla u. \tag{2.14}$$

Equation (2.13) states the balance between the heat flux p and the function f that represents the heat sources and sinks in the medium. Equation (2.14) is the constitutive relation between the heat flux and the temperature. The matrix coefficients are the heat conduction coefficients and they depend on the material.

This problem is known as the Poisson problem and it has several physical applications [49] besides the heat equation. We note that in all of them A describes the behavior of the medium occupying the domain Ω. Examples:

- Diffusion, where u is the concentration of the substance that diffuses in domain Ω, A represents the diffusion constants of the medium that define how the substance spreads to different directions at each point in the medium.
- Electrostatics, where u is the electrostatic potential, A is the dielectric constant and f is the charge density.
- Deflection of a membrane, where A describes the elastic properties of the medium. The general linear elasticity model is discussed below.

2.1.2 Linear elasticity

Let the domain $\Omega \subset \mathbb{R}^d$ have the boundary Γ consisting of two disjoint parts Γ_D and Γ_N. We assume that meas $\Gamma_D > 0$. The classical formulation of the linear elasticity problem is to find a tensor-valued function σ (stress) and a vector-valued function u (displacement) that satisfy the system of equations (see, e.g., [52])

$$\sigma = C\varepsilon \text{ in } \Omega, \tag{2.15}$$
$$\mathrm{Div}\sigma + f = 0 \quad \text{in } \Omega, \tag{2.16}$$
$$u = g \quad \text{on } \Gamma_D, \tag{2.17}$$
$$\sigma\nu = F \quad \text{on } \Gamma_N. \tag{2.18}$$

Here $g \in H^{1/2}(\Gamma_D)$ defines the Dirichlet boundary condition, f is the body force, and F is the traction on the boundary Γ_N. $C = \{c_{ijkm}\}$ is the tensor of elasticity

constants. The relation (2.15) is the *constitutive relation* (Hooke's law) defined by the medium and (2.16) is the *equation of equilibrium* between stresses and body forces.
A generalized solution of (2.15)-(2.18) is a function $u \in V_0 + g$, where

$$V_0 := \{w \in H^1(\Omega, \mathbb{R}^d) \mid w = 0 \quad \text{on } \Gamma_D\} \tag{2.19}$$

that minimizes the energy functional

$$J(w) := \tfrac{1}{2}\mathcal{E}(w) - \langle \ell, w \rangle,$$

where

$$\langle \ell, w \rangle = \int_\Omega f \cdot w \, dx + \int_{\Gamma_N} F \cdot w \, ds,$$

and

$$\mathcal{E}(u) = \int_\Omega \sigma(u) : \varepsilon(u) \, \mathrm{dx} = \int_\Omega C\varepsilon(u) : \varepsilon(u) \, \mathrm{dx} := \| \varepsilon(u) \|_C^2$$

is the elastic energy of the deformation. The generalized solution satisfies the integral relation

$$\int_\Omega C \, \varepsilon(u) : \varepsilon(w) \, dx = \langle \ell, w \rangle \quad \forall w \in V_0. \tag{2.20}$$

In the special case of an isotropic medium, the tensor C can be expressed with the help of only two material parameters. Usually, they are the Lame constants that lead to the form

$$C\varepsilon(u) = \lambda \mathrm{div} u \mathbb{I} + 2\mu\varepsilon(u). \tag{2.21}$$

Another pair of constants is Young's modulus E and Poisson's ratio ν. They are related to the Lame constants as follows:

$$\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}, \qquad \mu = \frac{E}{2(1+\nu)}. \tag{2.22}$$

2.1.3 Magnetostatics

One of the simplest forms of the equations used in electromagnetic problems is the following,

$$\begin{aligned} \mathrm{curl}(\mu^{-1}\mathrm{curl} u) + \kappa u &= f, \quad \text{in } \Omega, \\ u \times n &= 0, \quad \text{on } \Gamma. \end{aligned}$$

Here u is the vector valued function representing the electric field and μ and κ are physical parameters defined by the medium. The solution $u \in H_0(\Omega, \mathrm{curl})$ minimizes the energy functional

$$J(w) = \tfrac{1}{2} \int_\Omega \left(\mu^{-1} |\mathrm{curl} w|^2 + \kappa |w|^2 \right) \mathrm{dx} - \int_\Omega f \cdot w \, \mathrm{dx}$$

over the set

$$V_0 := \{w \in H(\Omega, \mathrm{curl}) \,|\, w \times n = 0\}.$$

The minimizer u satisfies the relation

$$\int\limits_{\Omega} \left(\mu^{-1}\mathrm{curl}u \cdot \mathrm{curl}w + \kappa u \cdot w\right) \,\mathrm{dx} = \int\limits_{\Omega} f \cdot w \,\mathrm{dx}, \quad \forall w \in V_0.$$

2.1.4 Stationary reaction diffusion problem

The reaction-diffusion problem with mixed Dirichlét–Robin boundary conditions is defined by the system

$$-\mathrm{div}(A\nabla u) + \rho u = f \quad \text{in } \Omega, \tag{2.23}$$
$$u = 0 \quad \text{on } \Gamma_D, \tag{2.24}$$
$$n \cdot A\nabla u = F \quad \text{on } \Gamma_N, \tag{2.25}$$
$$\alpha u + n \cdot A\nabla u = G \quad \text{on } \Gamma_R. \tag{2.26}$$

Here, $\Omega \subset \mathbb{R}^d$ has boundary $\Gamma_D \cup \Gamma_N \cup \Gamma_R$. Functions $A \in L_\infty(\Omega, \mathbb{M}^{d\times d})$ (symmetric and positive definite), $\rho \in L_\infty(\Omega, \mathbb{R}_+)$, and $\alpha \in L_\infty(\Gamma_R, \mathbb{R}_+)$ are related to the properties of the medium.

The reaction diffusion models are typically applied in chemistry, where they model how one or more substances distribute in the medium. The reaction term describes the local chemical reaction and the diffusion term causes substances to spread in the medium. In many models, the reaction term is non–linear with respect to the u. In the Thesis we study only linear models.

The energy functional for the problem is

$$J(w) = \frac{1}{2}\left(\int\limits_{\Omega} A\nabla w \cdot \nabla w \,\mathrm{dx} + \int\limits_{\Gamma_R} \alpha w^2 \,\mathrm{ds}\right) - \int\limits_{\Omega} fw \,\mathrm{dx} - \int\limits_{\Gamma_N} Fw \,\mathrm{ds} - \int\limits_{\Gamma_R} Gw \,\mathrm{ds}$$

and the space of admissible solutions is

$$V_0 := \{w \in H^1_0(\Omega) \,|\, w = 0 \text{ on } \Gamma_D\}. \tag{2.27}$$

The generalized solution is the function $u \in V_0$ satisfying the integral identity,

$$\int\limits_{\Omega} A\nabla u \cdot \nabla w \,\mathrm{dx} + \int\limits_{\Gamma_R} \alpha u w \,\mathrm{ds} = \int\limits_{\Omega} fw \,\mathrm{dx} - \int\limits_{\Gamma_N} Fw \,\mathrm{ds} - \int\limits_{\Gamma_R} Gw \,\mathrm{ds}, \;\; \forall w \in V_0.$$

2.2 General form of the variational problem

In this Section, we define a general form of the variational problem, which represents all physical models described in Section (2.1). The existence result proved for the abstract model is well known (see, e.g., [9]). The main motivation for the

TABLE 1 Definitions of the generalized model that correspond to different physical models.

	V_0	$\mathcal{A}$	Λ	$\mathcal{B}$
heat equation	$H_0^1(\Omega)$	A	∇	none
linear elasticity	V_0 (2.19)	C	ε	none
magnetostatics	$H_0(\Omega, \mathrm{curl})$	μ	curl	κ
reaction diffusion	V_0 (2.27)	A	∇	ρ

generalization presented here is to introduce a uniform notation that allows us to analyze an entire class of physical models by investigating a single model.

Let $\mathcal{V}$ and U be two Hilbert spaces. We denote their inner products by $(\cdot,\cdot)_{\mathcal{V}}$ and $(\cdot,\cdot)_U$ respectively. These inner products generate norms $\|\cdot\|_{\mathcal{V}}$ and $\|\cdot\|_U$. We define linear operators $\mathcal{A} : U \to U$ and $\mathcal{B} : \mathcal{V} \to \mathcal{V}$. The operators $\mathcal{A}$ and $\mathcal{B}$ are symmetric, i.e., $(\mathcal{A}y, z)_U = (y, \mathcal{A}z)_U$ for all $y, z \in U$ and similarly for $\mathcal{B}$.

In particular, we define a bounded linear operator $\Lambda : V \to U$, where $V \subset \mathcal{V}$ is a Hilbert space generated by the inner product $(w,v)_V := (w,v)_{\mathcal{V}} + (\Lambda w, \Lambda v)_U$. A set $V_0 \subset V$ is a convex, closed and non-empty subspace of V. We have $V_0 \subset V \subset \mathcal{V} \subset V_0^*$.

Using these definitions, the energy functionals generating variational problems presented in Section 2.1 can be written in a single abstract form. We define an energy functional $J : V \to \mathbb{R}$ as follows:

$$J(w) := \tfrac{1}{2}(\mathcal{A}\Lambda w, \Lambda w)_U + \tfrac{1}{2}(\mathcal{B}w, w)_{\mathcal{V}} - (f, w)_{\mathcal{V}}, \tag{2.28}$$

where $f \in \mathcal{V}$. In some of the energy functionals there is no term related to $\mathcal{B}$ and we have

$$J(w) := \tfrac{1}{2}(\mathcal{A}\Lambda w, \Lambda w)_U - (f, w)_{\mathcal{V}}. \tag{2.29}$$

In Table 1, we present how the generalized problem (2.28) can be defined to obtain the physical models discussed in Section 2.1. The set V_0 is defined as follows:

$$V_0 := \{w \in V \mid w \text{ satisfies homogeneous Dirichlet boundary conditions.}\}.$$

Note that in all physical models V is some Sobolev space associated to differential operator Λ. Spaces $\mathcal{V}$ and U are corresponding L_2–spaces of functions from the domain and the range of the differential operator respectively.

For convenience, we introduce the form $a : V_0 \times V_0 \to \mathbb{R}$,

$$a(u, w) := (\mathcal{A}\Lambda u, \Lambda w)_U + (\mathcal{B}u, w)_{\mathcal{V}}.$$

which is symmetric and bilinear. We assume that the bilinear form a is *continuous* in V_0, i.e., there exists $C > 0$ such that

$$a(u, v) \le C\|u\|_V \|w\|_V, \quad \forall u, v \in V_0$$

and *elliptic* in V_0, i.e., there exists $\alpha > 0$ such that

$$a(w,w) \geq \alpha \|w\|_V^2, \quad \forall w \in V_0. \tag{2.30}$$

The definition of a allows us to write energy functionals of the form (2.28) (and (2.29), where $\mathcal{B}$ is zero) as follows:

$$J(w) := \tfrac{1}{2} a(w,w) - (f,w)_{\mathcal{V}}. \tag{2.31}$$

The (generalized) solution u of the energy minimization (variational) problem

$$J(u) = \min_{w \in V_0} J(w), \tag{2.32}$$

has to satisfy the relation (weak form of the variational problem)

$$a(u,w) = (f,w)_{\mathcal{V}}, \quad \forall w \in V_0. \tag{2.33}$$

Note that for specific forms (2.28) and (2.29), the relation (2.33) reads as

$$(\mathcal{A}\Lambda u, \Lambda w)_U + (\mathcal{B}u, w)_{\mathcal{V}} = (f,w)_{\mathcal{V}}, \quad \forall w \in V_0 \tag{2.34}$$

and

$$(\mathcal{A}\Lambda u, \Lambda w)_U = (f,w)_{\mathcal{V}}, \quad \forall w \in V_0, \tag{2.35}$$

respectively.

It is easy to show that solutions of (2.32) and (2.33) coincide and there is utmost one solution. Indeed, for all $u, v \in V_0$ and $t \in \mathbb{R}$, we have

$$\begin{aligned} J(u+tw) &= \tfrac{1}{2} a(u+tw, u+tw) - (f, u+tw)_{\mathcal{V}} \\ &= J(u) + t\{a(u,w) - (f,w)_{\mathcal{V}}\} + \tfrac{1}{2} t^2 a(w,w). \end{aligned}$$

Now, if u satisfies (2.33), then by (2.30)

$$J(u+tv) = J(u) + \tfrac{1}{2} t^2 a(w,w) > J(u), \quad \forall w \in V_0,\ w \neq 0,$$

which shows that u is the minimizing functional. On the other hand, if u is the minimizer of J, then the Gateaux derivative of J at u must vanish, i.e.,

$$DJ(u) = a(u,w) - (f,w)_{\mathcal{V}} = 0, \quad \forall w \in V_0.$$

From ellipticity it follows that J is *coercive*, i.e.,

$$J(v_k) \to \infty, \quad \text{as} \quad \|v_k\|_V \to \infty.$$

Since J is convex, the existence and uniqueness can be easily proved (see, e.g., [9]).

Lemma 2.1. *Let J be coercive on V_0; then the set*

$$V_\gamma := \{v \in V_0 \mid J(v) \leq \gamma\}$$

is bounded in V_0.

Proof. Assume the opposite, i.e., that V_γ is unbounded and is not contained in any ball

$$B(0,d) = \{v \in V \mid \|v\|_V \leq d\}.$$

Then, for any integer k, one can find $v_k \in V_\gamma$ such that $\|v_k\|_V > k$. Then, by the coercivity we conclude that

$$J(v_k) \to \infty \quad \text{as} \quad k \to \infty.$$

This is a contradiction, because the elements of V_γ are such that the functional J does not exceed γ. □

Theorem 2.1. *Let $J : V_0 \to \mathbb{R}$ be convex, continuous and coercive. Let V_0 be a non-empty, convex and closed subset of Hilbert space V. Then the problem*

$$\inf_{w \in V_0} J(w)$$

has a minimizer u. If J is strictly convex, the minimizer is unique.

Proof. Let $\{v_k\}$ be a minimizing sequence, i.e., $J(v_k) \to \inf\limits_{V_0} \mathrm{J}$. Then the set

$$K_1 := \{v \in V_0 \mid J(v) \leq J(v_1)\}$$

is bounded (see Lemma 2.1). It is also closed. In a Hilbert space all closed bounded sets are weakly compact. Therefore, we can extract a weakly converging subsequence

$$\{v_{ks}\} \rightharpoonup u \in K_1.$$

Since J is convex and continuous, it is weakly lower semi continuous. We find that

$$\inf_{V_0} = \lim_{s \to \infty} J(v_{ks}) \geq J(u).$$

Hence u is the minimizer.

Assume that J is strictly convex, i.e.,

$$J(\lambda_1 v_1 + \lambda_2 v_2) < \lambda_1 J(v_1) + \lambda_2 J(v_2), \quad \lambda_1 + \lambda_2 = 1,\ \lambda_1, \lambda_2 > 0.$$

Assume that u_1 and u_2 are two different minimizers. Then we arrive to a contradiction because

$$J(\lambda_1 v_1 + \lambda_2 v_2) < \lambda_1 J(v_1) + \lambda_2 J(v_2) = \inf_{V_0} J.$$

□

The ellipticity of the bilinear form is crucial for the existence and uniqueness of the solution. The ellipticity of operators $\mathcal{A}$ and $\mathcal{B}$, i.e.,

$$(\mathcal{A}y, y)_U \geq c_1 \|y\|_U^2, \quad \forall\, y \in U, \tag{2.36}$$

and

$$(\mathcal{B}w, w)_{\mathcal{V}} \geq \tilde{c}_1 \|w\|_{\mathcal{V}}^2, \quad \forall\, w \in \mathcal{V}, \tag{2.37}$$

is often easy to establish. The inequality

$$\|w\|_{\mathcal{V}} \leq C_F \|\Lambda w\|_U, \quad \forall\, w \in V_0 \tag{2.38}$$

is the critical part. For the physical models presented in the Thesis, these embedding inequalities were briefly discussed on page 18. Note that by (2.36), (2.37), and (2.38) the bilinear forms related to energy functionals (2.28) and (2.29) are elliptic. Indeed, from (2.38) it follows that

$$\|w\|_V^2 = \|w\|_{\mathcal{V}}^2 + \|\,\Lambda w\|_U^2 \leq (C_F^2 + 1)\|\Lambda w\|_U^2, \quad \forall\, w \in V_0,$$

and we have

$$(\mathcal{A}\Lambda w, \Lambda w)_U \geq c_1 \|\Lambda w\|_U^2 \geq \frac{c_1}{C_F^2 + 1}\|w\|_V^2, \quad \forall\, w \in V_0$$

and

$$(\mathcal{A}\Lambda w, \Lambda w)_U + (\mathcal{B}w, w)_{\mathcal{V}} \geq c_1\|\Lambda w\|_U^2 + \tilde{c}_1\|\,w\|_{\mathcal{V}}^2 \geq \min\{c_1, \tilde{c}_1\}\|\,w\|_V^2, \quad \forall\, w \in V_0.$$

We define the adjoint operator $\Lambda^* : U \to V_0^*$, by the property

$$\langle \Lambda^* y, w \rangle = (y, \Lambda w)_U, \quad \forall\, y \in U,\ w \in V_0,$$

where $\langle \cdot, \cdot \rangle$ denotes the pairing of V_0 and its conjugate V_0^* and $\langle \Lambda^* y, w \rangle$ is the value of the functional $\Lambda^* y \in V_0^*$ at $w \in V_0$.

If we know a priori that $\Lambda^* \mathcal{A}\Lambda u \in \mathcal{V}$, then the weak form of the problem (2.35) can be rewritten as follows: find $u \in V_0$, such that

$$(\Lambda^* \mathcal{A}\Lambda u, w)_{\mathcal{V}} = (f, w)_{\mathcal{V}}, \quad \forall\, w \in V_0,$$

which yields

$$\Lambda^* A \Lambda u = f. \tag{2.39}$$

The boundary conditions for the problem are defined in the definition of the set V_0. This is the classical form (Euler equations) of the problem. It can be shown that if the classical form (2.39) has a solution, it coincides with the solution of the weak form (2.35). However, the weak form may have a solution that does not exist in the classical sense. Our analysis is based on the weak form of the problem.

Typically in physical models, $\mathcal{A}$ defines some physical properties of the medium occupying the domain Ω. An important quantity related to the solutions is the flux

$$p := \mathcal{A}\Lambda u.$$

Depending on the problem, it has different physical meanings. For example, in the linear elasticity problems it is the stress. We define a space of admissible fluxes,

$$Q := \{y \in U \mid \Lambda^* y \in \mathcal{V}\} \subset U.$$

For admissible fluxes, we can write

$$(\Lambda^* y, w)_{\mathcal{V}} = (y, \Lambda w)_U, \quad \forall\, y \in Q,\ w \in V_0. \tag{2.40}$$

Applying the definition of flux, (2.39) can be decomposed as

$$\begin{aligned} \Lambda^* p &= f && (2.41)\\ p &= \mathcal{A}\Lambda u. && (2.42) \end{aligned}$$

The relation (2.41) is referred as the *equilibrium relation* and (2.42) as the *constitutive relation*. They are the key relations defining the problem. The equilibrium relation is related to Newton's third law that states that for every action there is an equal and opposite reaction. A similar decomposition for the problem (2.34) is

$$\begin{aligned} \Lambda^* p + \mathcal{B}u &= f && (2.43)\\ p &= \mathcal{A}\Lambda u. && (2.44) \end{aligned}$$

The bilinear form defines an *energy norm* to the space V_0,

$$|\!|\!| \, w \, |\!|\!| := \sqrt{a(w,w)}.$$

Note that the definition of the norm depends on the particular problem. Since $\mathcal{A}$ is symmetric and positive definite, we can define additional norms:

$$\|y\|^2_{\mathcal{A}} := (\mathcal{A}y, y)_U$$

and

$$\|y\|^2_{\mathcal{A}^{-1}} := (\mathcal{A}^{-1}y, y)_U.$$

It is easy to show that

$$(y, q)_U \leq \| \, y\|_{\mathcal{A}} \| \, q\|_{\mathcal{A}^{-1}}, \quad \forall\, y, q \in U. \tag{2.45}$$

Indeed, since $\mathcal{A}$ is positive definite and symmetric, we have

$$0 \leq (\mathcal{A}(\mathcal{A}^{-1}y + \gamma q), \mathcal{A}^{-1}y + \gamma q)_U = (\mathcal{A}^{-1}y, y)_U - 2\gamma(y, q)_U + \gamma^2(\mathcal{A}q, q)_U,$$

where we can set $\gamma := \frac{(\mathcal{A}^{-1}y, y)_U}{(y,q)_U}$ (we assume that $y \neq 0$ and $(y, q)_U \neq 0$, otherwise (2.45) is trivially valid) and arrive at (2.45).

2.3 Estimates of deviations from exact solutions

The main tools for the analysis of the Thesis are the deviation estimates. First versions of these estimates were derived in [35] and consequent discussion and other related results can be found in [36, 37, 40, 44]. Full exposition of the related theory and collection of results can be found in the two books [32] and [39].

These estimates are derived by purely functional methods without any special information on approximations or numerical methods used. In this section, we derive deviation estimates for the problem of the form (2.35) and (2.34). On a practical and theoretical level, they provide a basis for various numerical error estimation and error indication schemes. Moreover, they can be applied to study the modelling error (see, e.g., [41, 42, 43]) and the error related to incompletely known data. The derivation methods presented here are applicable only to linear problems constructed by quadratic energy functionals. However, similar estimates can be derived for much more general problems by different methods (see, e.g., [39]). Henceforth we assume that the constants in (2.36) and (2.38) are either known or can be estimated.

We begin by deriving the deviation estimates for the following simple model problem:

$$\begin{aligned} -\Delta u &= f, \quad \text{in } \Omega, \\ u &= 0, \quad \text{on } \Gamma. \end{aligned}$$

The corresponding energy functional is

$$J(w) := \frac{1}{2} \int_\Omega |\nabla w|^2 \, \mathrm{dx} - \int_\Omega f w \, \mathrm{dx}.$$

The generalized solution $u \in H_0^1(\Omega)$ satisfies the integral relation

$$\int_\Omega \nabla u \cdot \nabla w \, \mathrm{dx} = \int_\Omega f w \, \mathrm{dx}, \quad \forall w \in H_0^1(\Omega). \tag{2.46}$$

The energy norm for the problem is

$$\| w \|^2 := \int_\Omega |\nabla w|^2 \, \mathrm{dx}.$$

Proposition 2.1. *Let $v \in H_0^1(\Omega)$ and u be the solution of the problem* (2.46). *Then we can estimate the deviation from both sides,*

$$\mathcal{M}_\ominus(v, w) \leq \| u - v \|^2 \leq \mathcal{M}_\oplus(v, y, \beta), \quad \forall w \in H_0^1(\Omega),\ y \in H(\Omega, \mathrm{div}),\ \beta > 0,$$

where

$$\mathcal{M}_\ominus(v, w) := -\int_\Omega \nabla w \cdot \nabla(w + 2v) \, dx + 2 \int_\Omega f w \, dx$$

and

$$\mathcal{M}_\oplus(v, y, \beta) := (1+\beta)\|\nabla v - y\|^2 + \frac{1+\beta}{\beta} c_F^2 \|\mathrm{div} y + f\|^2,$$

where c_F is from the inequality (2.8). *The inequalities are valid as equalities if $w := u - v$ and $y := \nabla u$.*

Proof. First, we construct the lower bound. By the definition of the energy norm and (2.46) we can write

$$\begin{aligned}
\|\!| u - v |\!\|^2 &= \int\limits_\Omega \nabla(u-v)\cdot\nabla(u-v)\,\mathrm{dx} \\
&= \|\!| u |\!\|^2 - 2\int\limits_\Omega \nabla u\cdot\nabla v\,\mathrm{dx} + \|\!| v |\!\|^2 + 2\left(-\|\!| u |\!\|^2 + \int\limits_\Omega fu\,\mathrm{dx}\right) \\
&= 2(J(v) - J(u)).
\end{aligned}$$

Since u is the minimizer of J, we can estimate from below $J(u) \leq J(v+w)$, where $w \in H_0^1(\Omega)$ is arbitrary. We arrive at

$$\begin{aligned}
\|\!| u - v |\!\|^2 &\geq 2(J(v) - J(v+w)) \\
&= \|\!| v |\!\|^2 - 2\int\limits_\Omega fv\,\mathrm{dx} - \|\!| v+w |\!\|^2 + 2\int\limits_\Omega f(v+w)\,\mathrm{dx} \\
&= -\|\!| w |\!\|^2 - 2\int\limits_\Omega \nabla v\cdot\nabla w\,\mathrm{dx} + 2\int\limits_\Omega fw\,\mathrm{dx} \\
&= -\int\limits_\Omega \nabla w\cdot\nabla(w+2v)\,\mathrm{dx} + 2\int\limits_\Omega fw\,\mathrm{dx}.
\end{aligned}$$

Clearly, the lower bound is sharp if $w := u - v$.

Next, we construct the upper bound. From (2.46) follows:

$$\int\limits_\Omega \nabla(u-v)\cdot\nabla w\,\mathrm{dx} = -\int\limits_\Omega \nabla v\cdot\nabla w\,\mathrm{dx} + \int\limits_\Omega fw\,\mathrm{dx} + \int\limits_\Omega (\mathrm{div} y w + y\cdot\nabla w)\,\mathrm{dx}.$$

Due to (2.4), the last term is zero for any $y \in H(\Omega, \mathrm{div})$ and for all $w \in H_0^1(\Omega)$. We can write

$$\int\limits_\Omega \nabla(u-v)\cdot\nabla w\,\mathrm{dx} = \int\limits_\Omega (y - \nabla v)\cdot\nabla w\,\mathrm{dx} + \int\limits_\Omega (\mathrm{div} y + f)w\,\mathrm{dx}.$$

We can estimate from above by (2.5),

$$\int\limits_\Omega \nabla(u-v)\cdot\nabla w\,\mathrm{dx} \leq \|y - \nabla v\| \ \|\!| w |\!\| + \|\mathrm{div} y + f\|\,\|w\|.$$

By (2.8),

$$\int\limits_\Omega \nabla(u-v)\cdot\nabla w\,\mathrm{dx} \leq (\|y - \nabla v\| + c_F\|\mathrm{div} y + f\|)\ \|\!| w |\!\|,$$

substituting $w := u - v$ leads to,

$$\|\!| u - v |\!\| \leq \|y - \nabla v\| + c_F\|\mathrm{div} y + f\|.$$

Squaring both sides and using (2.10) leads to the stated upper bound. If $y = \nabla u$ the second part of the majorant vanishes and the first one provides the exact error. □

A lower deviation estimate for the generalized problem (2.33)

We construct a lower bound for the deviation from the exact solution. Theorem 2.2 presented below is valid for any problem of the form (2.33).

Theorem 2.2. *Let u be the solution of* (2.33) *and $v \in V_0$. Then for the functional*

$$\mathcal{M}_{\ominus}(v,w) := -a(w, w+2v) + 2(f,w)_{\mathcal{V}} \tag{2.47}$$

it follows that

$$|\!|\!| \; u - v \; |\!|\!|^2 \geq \mathcal{M}_{\ominus}(v,w), \quad \forall w \in V_0$$

and

$$|\!|\!| \; u - v \; |\!|\!|^2 = \sup_{w \in V_0} \mathcal{M}_{\ominus}(v,w).$$

Proof. Let u be the solution of (2.33) and $v \in V_0$. By the definition of the energy norm,

$$\begin{aligned} |\!|\!| \; u - v \; |\!|\!|^2 &= a(u-v, u-v) \\ &= a(u,u) - 2a(u,v) + a(v,v) + 2((f,u)_{\mathcal{V}} - a(u,u)) \\ &= 2\left(\tfrac{1}{2}a(v,v) - (f,v)_{\mathcal{V}} - \tfrac{1}{2}a(u,u) + (f,u)_{\mathcal{V}}\right) \\ &= 2(J(v) - J(u)), \end{aligned}$$

since u satisfies the Galerkin orthogonality condition. By definition $J(u) \leq J(w)$, for all $w \in V_0$. We substitute $u := w + v$ and estimate from below

$$\begin{aligned} |\!|\!| \; u - v \; |\!|\!|^2 &\geq 2\left(\tfrac{1}{2}a(v,v) - (f,v)_{\mathcal{V}} - \tfrac{1}{2}a(w+v, w+v) + (f, w+v)_{\mathcal{V}}\right) \\ &= -a(w, w+2v) + 2(f,w)_{\mathcal{V}}. \end{aligned}$$

The inequality becomes an equality if $w := u - v$. □

Henceforth we refer to the functional $\mathcal{M}_{\ominus}$ as the minorant.

An upper deviation estimate for the problem of type (2.35)

The upper deviation estimate related to the problem (2.35) is defined as follows:

$$\mathcal{M}_{\oplus}(v,y,\beta) := (1+\beta)(\Lambda v - \mathcal{A}^{-1}y, \mathcal{A}\Lambda v - y)_U + \left(1 + \frac{1}{\beta}\right)\frac{C_F^2}{c_1}\|f - \Lambda^* y\|_{\mathcal{V}}^2, \tag{2.48}$$

where C_F and c_1 are from (2.38) and (2.36) respectively. This functional (2.48) we will call the majorant and it is the natural counterpart of the minorant defined in (2.47). The structure of the majorant shows that it is related to the violations of the physical laws (2.41) and (2.42). The energy norm for problem (2.35) is

$$|\!|\!| \; w \; |\!|\!|^2 := (\mathcal{A}\Lambda w, \Lambda w)_U.$$

Theorem 2.3. *Let u be the solution of* (2.35) *and $v \in V_0$. Then,*

$$|\!|\!| u - v |\!|\!|^2 \leq \mathcal{M}_{\oplus}(v, y, \beta), \quad \forall y \in Q \text{ and } \beta > 0,$$

and

$$|\!|\!| u - v |\!|\!|^2 = \inf_{\substack{y \in Q, \\ \beta > 0}} \mathcal{M}_{\oplus}(v, y, \beta).$$

Proof. By subtracting $(A\Lambda v, \Lambda w)_U$ from (2.35) and applying (2.40) yields

$$\begin{aligned}(\mathcal{A}\Lambda(u - v), \Lambda w)_U &= (f, w)_{\mathcal{V}} - (\mathcal{A}\Lambda v, \Lambda w)_U - (\Lambda^* y, w)_{\mathcal{V}} + (y, \Lambda w)_U \\ &= (y - \mathcal{A}\Lambda v, \Lambda w)_U + (-\Lambda^* y + f, w)_{\mathcal{V}},\end{aligned}$$

where $y \in Q$ is arbitrary. We estimate the first term from above by (2.45) and the second term by the Cauchy-Schwartz inequality, (2.36) and (2.38),

$$\begin{aligned}(\mathcal{A}\Lambda(u - v), \Lambda w)_U &\leq (\mathcal{A}^{-1}(y - \mathcal{A}\Lambda v), y - \mathcal{A}\Lambda v)_U^{\frac{1}{2}} (\mathcal{A}\Lambda w, \Lambda w)_U^{\frac{1}{2}} + \\ &\quad + \| - \Lambda^* y + f\|_{\mathcal{V}} \|w\|_{\mathcal{V}} \\ &\leq (\mathcal{A}^{-1}(y - \mathcal{A}\Lambda v), y - \mathcal{A}\Lambda v)_U^{\frac{1}{2}} |\!|\!| w |\!|\!| + \\ &\quad + \frac{C_F}{\sqrt{c_1}} \| - \Lambda^* y + f\|_{\mathcal{V}} |\!|\!| w |\!|\!| .\end{aligned}$$

Since w is arbitrary, we can substitute $w := u - v$ and obtain $|\!|\!| u - v |\!|\!|^2$ on the left-hand side and $|\!|\!| u - v |\!|\!|$ as a common multiplier on the right-hand side. Dividing by the energy norm leads to

$$|\!|\!| u - v |\!|\!| \leq (\Lambda v - \mathcal{A}^{-1} y, \mathcal{A}\Lambda v - y)_U^{\frac{1}{2}} + \frac{C_F}{\sqrt{c_1}} \|f - \Lambda^* y\|_{\mathcal{V}}.$$

As before, we square both sides and apply (2.10) and obtain (2.48). Note that if we set $y := \mathcal{A}\Lambda u$ the inequality becomes an equality. □

An upper deviation estimate for the generalized problem of type (2.34)

Although the problem (2.34) has a form similar to problem (2.35), the presence of the operator $\mathcal{B}$ provides more alternatives for the derivation of the majorant. (we note that the minorant, nevertheless, has the same form as defined in Theorem 2.2). As an introductory example, we derive three different majorants for a simple reaction diffusion type model problem and discuss their properties. Then we repeat a similar derivation for the more complicated reaction diffusion model (2.23–2.26) and the generalized problem (2.34).

Consider the problem:

$$\begin{aligned}-\Delta u + \rho u &= f, \quad \text{in } \Omega \\ u &= 0, \quad \text{on } \Gamma,\end{aligned}$$

where $\rho \in L_\infty(\Omega, \mathbb{R}_+)$. The generalized solution $u \in H^1_0(\Omega)$ satisfies the integral identity

$$\int_\Omega (\nabla u \cdot \nabla w + \rho u w)\,\mathrm{dx} = \int_\Omega f w\,\mathrm{dx}, \quad \forall w \in H^1_0(\Omega). \tag{2.49}$$

Note that the problem can be decomposed to constitutive and equilibrium relations,

$$-\mathrm{div} p + \rho u = f, \tag{2.50}$$

$$p = \nabla u, \tag{2.51}$$

where p is the exact flux. For notational purposes we denote the violation of the relation (2.50) by the residual function,

$$r(v, y) := f + \mathrm{div} y - \rho v. \tag{2.52}$$

The problem has the energy norm

$$||| \, w \, |||^2 := \int_\Omega \left(|\nabla w|^2 + \rho w^2 \right) \mathrm{dx}.$$

First, we present two different upper deviation estimates, which each have their benefits and drawbacks.

Proposition 2.2. *Let $v \in H^1_0(\Omega)$ and u be the solution of* (2.49). *Then*

$$||| \, u - v \, |||^2 \leq \|\tfrac{1}{\sqrt{\rho}} r(v, y)\|^2 + \|y - \nabla v\|^2, \quad \forall y \in H(\Omega, \mathrm{div}). \tag{2.53}$$

Proof. We subtract $\int_\Omega (\nabla v \cdot \nabla w + \rho v w)\,\mathrm{dx}$ from (2.49) and apply (2.4) to obtain

$$\begin{aligned}
\int_\Omega (\nabla(u - v) \cdot \nabla w + \rho(u - v) w)\,\mathrm{dx} &= \int_\Omega (f w - \nabla v \cdot \nabla w - \rho v w + \mathrm{div} y w + y \cdot \nabla w)\,\mathrm{dx} \\
&= \int_\Omega (f + \mathrm{div} y - \rho v) w\,\mathrm{dx} + \int_\Omega (y - \nabla v) \cdot \nabla w\,\mathrm{dx} \\
&= \int_\Omega \tfrac{1}{\sqrt{\rho}} r(v, y) \sqrt{\rho} w\,\mathrm{dx} + \int_\Omega (y - \nabla v) \cdot \nabla w\,\mathrm{dx}.
\end{aligned}$$

We estimate by (2.5),

$$\begin{aligned}
\int_\Omega (\nabla(u - v) \cdot \nabla w + \rho(u - v) w)\,\mathrm{dx} &\leq \|\tfrac{1}{\sqrt{\rho}} r(v, y)\| \, \|\sqrt{\rho} w\| + \|y - \nabla v\| \, \|\nabla w\| \\
&\leq \left(\|\tfrac{1}{\sqrt{\rho}} r(v, y)\|^2 + \|y - \nabla v\|^2 \right)^{\frac{1}{2}} \left(\|\sqrt{\rho} w\|^2 + \|\nabla w\|^2 \right)^{\frac{1}{2}}.
\end{aligned}$$

On the left-hand side, we constructed the energy norm of w as a common multiplier. Substituting $w := u - v$ leads at

$$||| \, u - v \, ||| \leq \left(\|\tfrac{1}{\sqrt{\rho}} r(v, y)\|^2 + \|y - \nabla v\|^2 \right)^{\frac{1}{2}}.$$

□

Proposition 2.3. *Let $v \in H_0^1(\Omega)$ and u be the solution of* (2.49). *Then*

$$\| u - v \|^2 \leq \frac{1+\beta}{\beta} c_F^2 \|r(v,y)\|^2 + (1+\beta)\|y - \nabla w\|^2, \quad \forall y \in H(\Omega, \mathrm{div}),\ \beta \in \mathbb{R}_+. \tag{2.54}$$

Proof. As in the proof of Proposition 2.2 we arrive at

$$\int_\Omega (\nabla(u-v) \cdot \nabla w + \rho(u-v)w)\,\mathrm{d}x \leq \int_\Omega r(v,y)w\,\mathrm{d}x + \int_\Omega (y - \nabla v) \cdot \nabla w\,\mathrm{d}x.$$

We estimate by (2.5) and (2.8) and arrive at

$$\begin{aligned} \int_\Omega (\nabla(u-v) \cdot \nabla w + \rho(u-v)w)\,\mathrm{d}x &\leq \|r(v,y)\|\,\|w\| + \|y - \nabla w\|\,\|\nabla w\| \\ &\leq (c_F\|r(v,y)\| + \|y - \nabla w\|)\,\|\nabla w\|. \end{aligned}$$

Clearly, $\|\nabla w\| \leq \| w \|$. Again, substituting $w := u - v$ leads to the estimate

$$\| u - v \| \leq c_F\|r(v,y)\| + \|y - \nabla w\|.$$

Squaring both sides and (2.10) leads to (2.54). □

Both majorants in Propositions 2.2 and 2.3 are clearly related to the violation of physical laws (2.50) and (2.51), but their applicability differs significantly.

The estimate in Proposition 2.2 is sharp, i.e., if we substitute $y := \nabla u$, the inequality becomes an equality. However, in practice exact flux is rarely at our disposal, so we operate with some approximations of it. If ρ is very small, then even small violations of the condition (2.52) cause the estimate to became overly pessimistic. The estimate in Proposition 2.3 does not suffer from this drawback, but it does not provide the exact deviation with any auxiliary flux y.

The solution is to split the residual term. This approach was used in [40] for the reaction-diffusion problem and in [44] for the generalized Stokes problem. In these publications it was shown that splitting of the residual term (performed with the help of a function μ) allows obtaining error majorants that are insensitive with respect to small values of the lower term coefficient and at the same time sharp (i.e., have no irremovable gaps between the left- and right-hand sides).

Proposition 2.4. *Let $v \in H_0^1(\Omega)$ and u be the solution of* (2.49). *Then*

$$\| u - v \|^2 \leq \|\tfrac{1}{\sqrt{\rho}}\mu r(v,y)\|^2 + \frac{1+\beta}{\beta} c_F^2 \|(1-\mu) r(v,y)\|^2 + (1+\beta)\|y - \nabla v\|^2, \tag{2.55}$$

$\forall y \in H(\Omega, \mathrm{div})$, $\beta > 0$, *where $\mu : \Omega \to [0,1]$ is arbitrary and c_F is from* (2.38).

Proof. As in the proof of Proposition 2.2 we arrive at

$$\int_\Omega (\nabla(u-v) \cdot \nabla w + \rho(u-v)w)\,\mathrm{d}x \leq \int_\Omega r(v,y)w\,\mathrm{d}x + \int_\Omega (y - \nabla v) \cdot \nabla w\,\mathrm{d}x.$$

We introduce the splitting function $\mu : \Omega \to [0,1]$, which is at our disposal. Then we decompose

$$\int_\Omega (\nabla(u-v)\cdot\nabla w + \rho(u-v)w)\,\mathrm{dx} \le \int_\Omega \mu r(v,y) w\,\mathrm{dx} + \\ + \int_\Omega (1-\mu) r(v,y) w\,\mathrm{dx} + \int_\Omega (y-\nabla v)\cdot\nabla w\,\mathrm{dx}.$$

We estimate the first term as in Proposition 2.2, by (2.5) and the second one as in Proposition 2.3 by (2.5) and (2.8),

$$\int_\Omega (\nabla(u-v)\cdot\nabla w + \rho(u-v)w)\,\mathrm{dx} \le \|\tfrac{1}{\sqrt{\rho}}\mu r(v,y)\|\,\|\rho w\| + \\ + c_F\|(1-\mu)r(v,y)\|\|\nabla w\| + \|y-\nabla v\|\,\|\nabla w\| \\ \le \left(\|\tfrac{1}{\sqrt{\rho}}\mu r(v,y)\|^2 + (c_F\|(1-\mu)r(v,y)\| + \|y-\nabla v\|)^2\right)^{\frac{1}{2}} \left(\|\rho w\|^2 + \|\nabla w\|^2\right)^{\frac{1}{2}}.$$

Substituting $w := u - v$ leads to,

$$|\!|\!|\; u-v \;|\!|\!| \le \left(\|\tfrac{1}{\sqrt{\rho}}\mu r(v,y)\|^2 + (c_F\|(1-\mu)r(v,y)\| + \|y-\nabla v\|)^2\right)^{\frac{1}{2}}.$$

Squaring both sides and using again (2.10) leads to (2.55). □

Note that in (2.55), if we set $y := \nabla u$ and $\mu = 1$, we obtain the exact error. It is easy to see that lowest value of (2.55) is obtained with μ that pointwise minimizes the quantity

$$\frac{1}{\rho}\mu^2 + \frac{1+\beta}{\beta}c_F(1-\mu)^2.$$

In [2], we derived upper deviation estimates of the type presented in Propositions 2.2, 2.3, and 2.4 for the magnetostatical problem defined in Section 2.1.3. We also included numerical tests to study the performance of all three different types of majorants. These tests were programmed by Anjam and Muzalevski using Nedelec type elements. Numerical tests by Nedelec elements for non–splitted majorant were first done in [15].

An upper deviation estimate for the reaction diffusion model (2.23–2.26)

Similar estimates can be derived for the more complicated reaction diffusion problem (2.23–2.26). The problem can be decomposed to relations

$$-\mathrm{div} p + \rho u = f \quad \text{in } \Omega, \tag{2.56}$$

$$A\nabla u = p \quad \text{in } \Omega, \tag{2.57}$$

$$\alpha u + n\cdot p = G \quad \text{on } \Gamma_R, \tag{2.58}$$

$$n\cdot p = F \quad \text{on } \Gamma_N, \tag{2.59}$$

$$u = 0 \quad \text{on } \Gamma_D. \tag{2.60}$$

We recall the weak form of the problem: Find $u \in V_0$ satisfying

$$\int\limits_{\Omega} (A\nabla u \cdot \nabla w + \rho u w)\ \mathrm{dx} + \int\limits_{\Gamma_R} \alpha u w \,\mathrm{ds} = \int\limits_{\Omega} f w \,\mathrm{dx} - \int\limits_{\Gamma_N} F w \,\mathrm{ds} - \int\limits_{\Gamma_R} G w \,\mathrm{ds}, \quad \forall w \in V_0.$$

Note that the bilinear form is defined as

$$a(u,w) := \int\limits_{\Omega} (A\nabla u \cdot \nabla w + \rho u w)\ \mathrm{dx} + \int\limits_{\Gamma_R} \alpha u w \,\mathrm{ds}.$$

Let $v \in V_0$ be an admissible approximation of the exact solution u (generated by A, ϱ, and α). From the weak form of the problem it follows that for any $w \in V_0$

$$\begin{aligned} a(u-v,w) = \int\limits_{\Omega} f w \,\mathrm{dx} + \int\limits_{\Gamma_N} F w \,\mathrm{ds} + \int\limits_{\Gamma_R} G w \,\mathrm{ds} - \\ - \int\limits_{\Omega} A\nabla v \cdot \nabla w \,\mathrm{dx} - \int\limits_{\Omega} \rho v w \,\mathrm{dx} - \int\limits_{\Gamma_R} \alpha v w \,\mathrm{ds} + \\ + \int\limits_{\Omega} (\mathrm{div}(y) w + y \cdot \nabla w)\ \mathrm{dx} - \int\limits_{\Gamma_N \cup \Gamma_R} (y \cdot \nu)\, w \,\mathrm{ds}, \end{aligned} \tag{2.61}$$

where ν denotes an unit outward normal to Γ and

$$y \in H^+(\Omega, \mathrm{div}) := \{ y \in H(\Omega, \mathrm{div}) \mid y \cdot \nu \in L^2(\Gamma_N \cup \Gamma_R) \}.$$

We note that the last line is zero for all $y \in H^+(\Omega, \mathrm{div})$ (in view of (2.4)). We regroup the terms and rewrite the relation as follows:

$$a(u-v,w) = I_1 + I_2 + I_3 + I_4, \tag{2.62}$$

where

$$\begin{aligned} I_1 &:= \int\limits_{\Omega} r_1(v,y) w \,\mathrm{dx} := \int\limits_{\Omega} (f - \rho v + \mathrm{div}(y)) w \,\mathrm{dx}, \\ I_2 &:= \int\limits_{\Gamma_R} r_2(v,y) w \,\mathrm{dx} := \int\limits_{\Gamma_R} (G - \alpha v - y \cdot \nu) w \,\mathrm{ds}, \\ I_3 &:= \int\limits_{\Gamma_N} (F - y \cdot \nu) w \,\mathrm{ds}, \\ I_4 &:= \int\limits_{\Omega} (y - A\nabla v) \cdot \nabla w \,\mathrm{dx}. \end{aligned}$$

Note that each term is related to a specific relation in (2.56-2.59). Now we can estimate each term separately by the Friedrichs and trace inequalities (which are valid due to our assumptions concerning Ω). We have

$$\begin{aligned} ||w||^2_{\Omega} &\leq C_1(\Omega) ||\nabla w||^2_{\Omega} \quad \forall w \in V_0, \\ ||w||^2_{2,\Gamma_N} &\leq C_2(\Omega, \Gamma_N) ||\nabla w||^2_{\Omega} \quad \forall w \in V_0, \\ ||w||^2_{2,\Gamma_R} &\leq C_3(\Omega, \Gamma_R) ||\nabla w||^2_{\Omega} \quad \forall w \in V_0. \end{aligned}$$

When estimating the integrands of I_1 and I_2, we introduce additional functions μ_1 and μ_2, which have values in $[0,1]$.

$$\begin{aligned} I_1 &= \int\limits_\Omega \tfrac{\mu_1}{\sqrt{\rho}} r_1(v,y)\sqrt{\rho} w \,\mathrm{dx} + \int\limits_\Omega (1-\mu_1) r_1(v,y) w \,\mathrm{dx} \\ &\leq \left\| \tfrac{\mu_1}{\sqrt{\rho}} r_1(v,y) \right\|_\Omega ||\sqrt{\rho} w||_\Omega + \sigma_1 ||(1-\mu_1) r_1(v,y)||_\Omega \left(\int\limits_\Omega A\nabla w \cdot \nabla w \, dx \right)^{\frac{1}{2}} \end{aligned}$$

and

$$\begin{aligned} I_2 &= \int\limits_{\Gamma_R} \tfrac{\mu_2}{\sqrt{\alpha}} r_2(v,y)\sqrt{\alpha} w \,\mathrm{dx} + \int\limits_{\Gamma_R} (1-\mu_2) r_2(v,y) w \,\mathrm{dx} \\ &\leq \left\| \tfrac{\mu_2}{\sqrt{\alpha}} r_2(v,y) \right\|_{\Gamma_R} ||\sqrt{\alpha} w||_{\Gamma_R} + \\ &\quad + ||(1-\mu_2) r_2(v,y)||_{\Gamma_R} \, \sigma_3 \left(\int\limits_\Omega A\nabla w \cdot \nabla w \, dx \right)^{\frac{1}{2}}, \end{aligned}$$

and

$$I_3 \leq ||F - y \cdot \nu \,||_{\Gamma_N} \sigma_2 \left(\int\limits_\Omega A\nabla w \cdot \nabla w \, dx \right)^{\frac{1}{2}}, \tag{2.63}$$

where

$$\sigma_1 = \sqrt{\frac{C_1(\Omega)}{c_1}}, \quad \sigma_2 = \sqrt{\frac{C_2(\Omega,\Gamma_N)}{c_1}}, \quad \text{and} \quad \sigma_3 = \sqrt{\frac{C_3(\Omega,\Gamma_3)}{c_1}}.$$

In our case we have two lower terms so that we need two functions μ_1 and μ_2 to split the respective residual terms.

The term I_4 is estimated (by (2.45) for $U := L_2(\mathbb{R}^d, \Omega)$) as follows:

$$I_4 \leq D(\nabla v, y)^{\frac{1}{2}} \left(\int\limits_\Omega A\nabla w \cdot \nabla w \, dx \right)^{\frac{1}{2}}, \tag{2.64}$$

where

$$D(\nabla v, y) = \int\limits_\Omega (y - A\nabla v) \cdot (\nabla v - A^{-1} y) \,\mathrm{dx}. \tag{2.65}$$

We collect all the terms and obtain

$$\begin{aligned} a(u-v,w) \leq \Big(& D(\nabla v,y)^{\frac{1}{2}} + \sigma_1 ||(1-\mu_1) r_1(v,y)||_\Omega + \\ & + \sigma_3 ||(1-\mu_2) r_2(v,y)||_{\Gamma_R} + \sigma_2 ||F - y\cdot\nu\,||_{\Gamma_N} \Big) \left(\int\limits_\Omega A\nabla w \cdot \nabla w \, dx \right)^{\frac{1}{2}} + \\ & + \left\| \tfrac{\mu_1}{\sqrt{\rho}} r_1(v,y) \right\|_\Omega ||\sqrt{\rho} w||_\Omega + \left\| \tfrac{\mu_2}{\sqrt{\alpha}} r_2(v,y) \right\|_{\Gamma_R} ||\sqrt{\alpha} w||_{\Gamma_R}. \end{aligned}$$

Set $w = u - v$ and use the Cauchy-Schwartz inequality. Then, we arrive at the estimate

$$\|\!| u - v \|\!|^2 \leq \Big(D(\nabla v, y)^{\frac{1}{2}} + \sigma_1 ||(1-\mu_1) r_1(v,y)||_\Omega + \\ + \sigma_3 ||(1-\mu_2) r_2(v,y)||_{\Gamma_R} + \sigma_2 ||F - y \cdot \nu ||_{\Gamma_N} \Big)^2 + \\ + \left\| \tfrac{\mu_1}{\sqrt{\rho}} r_1(v,y) \right\|^2_\Omega + \left\| \tfrac{\mu_2}{\sqrt{\alpha}} r_2(v,y) \right\|^2_{\Gamma_R}. \quad (2.66)$$

It is worth remarking that the estimate (2.66) provides a guaranteed upper bound of the error for *any* conforming approximation of the problem (2.23–2.26). The estimate has a form typical for all functional a posteriori estimates: it is the sum of residuals of the basic relations with multipliers that depend on the constants in the respective functional (embedding) inequalities for the domain and boundary parts.

However, for our subsequent goals, it is desirable to have the majorant in a form that involves quadratic terms only. Such a form can easily be derived from (2.66) if we square both parts and apply the Cauchy–Schwartz inequality again to the first term (with multipliers $\gamma_i > 0$, $i = 1,2,3,4$). Then we obtain

$$\|\!| u - v \|\!|^2 \leq \kappa \Big(\gamma_1 D(\nabla v, y) + \gamma_2 ||(1-\mu_1) r_1(v,y)||^2_\Omega + \\ + \gamma_3 ||(1-\mu_2) r_2(v,y)||^2_{\Gamma_R} + \gamma_4 ||F - y \cdot \nu ||^2_{\Gamma_N} \Big) + \\ + \left\| \tfrac{\mu_1}{\sqrt{\rho}} r_1(v,y) \right\|^2_\Omega + \left\| \tfrac{\mu_2}{\sqrt{\alpha}} r_2(v,y) \right\|^2_{\Gamma_R}. \quad (2.67)$$

where

$$\kappa := \frac{1}{\gamma_1} + \frac{\sigma_1^2}{\gamma_2} + \frac{\sigma_2^2}{\gamma_3} + \frac{\sigma_3^2}{\gamma_4}.$$

We note that (2.67) coincides with (2.66) if

$$\begin{aligned} \gamma_1 = \bar{\gamma}_1 &:= D(\nabla v, y)^{-\frac{1}{2}}, \\ \gamma_2 = \bar{\gamma}_2 &:= \frac{\sigma_1}{||(1-\mu_1) r_1(v,y)||_\Omega}, \\ \gamma_3 = \bar{\gamma}_3 &:= \frac{\sigma_2}{||(1-\mu_2) r_2(v,y)||_{\Gamma_R}}, \\ \gamma_4 = \bar{\gamma}_4 &:= \frac{\sigma_3}{||F - y \cdot \nu ||_{\Gamma_N}}. \end{aligned}$$

Certainly, the estimate (2.67) looks more complicated compared to (2.66). However, it has an important advantage: the weight functions μ_1 and μ_2 enter it as quadratic integrands, so that we can easily find their optimal form adapted to a particular v and the respective error distribution.

In the simplest case, we take $\mu_1 = \mu_2 = 0$, which yields the estimate

$$\|\!| u - v \|\!|^2 \leq \kappa \times \Big(\gamma_1 D(\nabla v, y) + \gamma_2 ||r_1(v,y)||^2_\Omega + \\ + \gamma_3 ||r_2(v,y)||^2_{\Gamma_R} + \gamma_4 ||F - y \cdot \nu ||^2_{\Gamma_N} \Big). \quad (2.68)$$

Another estimate arises if we set $\mu_1 = \mu_2 = 1$. In this case, the terms with factors σ_1 and σ_3 in (2.66) are equal to zero, so that subsequent relations do not contain the terms with multipliers formed by γ_2 and γ_3. Hence, we arrive at the estimate

$$\||\, u - v \,\||^2 \leq \left(\frac{1}{\gamma_1} + \frac{\sigma_3^2}{\gamma_4}\right) \times \left(\gamma_1 D(\nabla v, y) + \gamma_4 ||F - y \cdot \nu\,||^2_{\Gamma_N}\right) + \\ + \left\|\tfrac{1}{\sqrt{\rho}} r_1(v,y)\right\|^2_\Omega + \left\|\tfrac{1}{\sqrt{\alpha}} r_2(v,y)\right\|^2_{\Gamma_R}. \quad (2.69)$$

The estimate (2.69) involves "free" parameters γ_1 and γ_4 and a "free" vector–valued function y (which can be thought of as an image of the true flux). There exists a combination of these free parameters that makes the left-hand side of the estimate equal to the right-hand one. Indeed, set $y = A\nabla u$. Then

$$\begin{aligned} r_1(v,y) &= \rho(u - v) \quad \text{in } \Omega, \\ r_2(v,y) &= \alpha(u - v) \quad \text{on } \Gamma_R, \\ F - y \cdot \nu &= 0 \quad \text{on } \Gamma_N \end{aligned}$$

and we find that for γ_4 tending to infinity and for any $\gamma_1 > 0$ the right-hand side of (2.69) coincides with the energy norm of the error. However, the estimate (2.69) has a drawback: it is sensitive to ρ and α and may essentially overestimate the error if ρ or α are small. On the other hand, the right-hand side of (2.68) is stable with respect to small values of ρ and α. Regrettably, it does not possess the "exactness" (in the sense discussed above) because it may have a "gap" between the left- and right-hand sides for any y.

An upper bound of the error that combines the positive features of (2.68) and (2.69) can be derived (as in [40, 44]) if a certain optimization procedure is used to select optimal functions μ_1 and μ_2. If y is given, then optimal μ_1 and μ_2 can be found analytically. It is not difficult to see that μ_1 must minimize the quantity

$$\int\limits_\Omega \left(\kappa\gamma_2(1 - \mu_1)^2 + \tfrac{\mu_1^2}{\rho}\right) r_1(v,y)^2 dx.$$

The quantity attains its minimum with

$$\mu_1(x) = \mu_1^{opt}(x) := \frac{\kappa\gamma_2}{\kappa\gamma_2 + \rho(x)^{-1}} \quad \text{in } \Omega.$$

Similarly, we find that the integrals associated with Γ_R attain a minimum if

$$\mu_2(x) = \mu_2^{opt}(x) := \frac{\kappa\gamma_3}{\kappa\gamma_3 + \alpha(x)^{-1}} \quad \text{on } \Gamma_3.$$

Substituting these values to (2.67) results in the estimate

$$\||\, u - v \,\||^2 \leq \kappa\Big(\gamma_1 D(\nabla v, y) + \gamma_2 \left\|\tfrac{\sqrt{\kappa^2\gamma_2^2\rho + 1}}{\kappa\gamma_2\rho + 1}\, r_1(v,y)\right\|^2_\Omega + \\ + \gamma_3 \left\|\tfrac{\sqrt{\kappa^2\gamma_3^2\alpha + 1}}{\kappa\gamma_3\alpha + 1}\, r_2(v,y)\right\|^2_{\Gamma_R} + \gamma_4 ||F - y \cdot \nu\,||^2_{\Gamma_N}\Big). \quad (2.70)$$

We denote the right hand side of (2.70) by $\mathcal{M}_\oplus(v, y, \gamma, \mu_1, \mu_2)$. This error majorant provides a guaranteed upper bound of the error, i.e.,

$$|||\ u - v\ |||^2 \leq \mathcal{M}_\oplus(v, y, \gamma, \mu_1, \mu_2). \tag{2.71}$$

It is exact (in the sense discussed above). Indeed, for $y = A\nabla u$ and $\mu_1 = \mu_2 = 1$ we obtain

$$\inf_{\gamma_i > 0} \mathcal{M}_\oplus(v, A\nabla u, \gamma, 1, 1) = |||\ u - v\ |||^2 . \tag{2.72}$$

Also, it directly follows from the structure of (2.70) that the right-hand side is insensitive to small values of ρ and α.

An upper deviation estimate for the generalized model (2.34)

We can derive a similar upper bound for the general problem (2.34) using the same methods. Again, we introduce a weight function $\mu : \mathcal{V} \to \mathcal{V}$, which is multiplication by a scalar number that attains values from 0 to 1. We denote the class of these multiplication functions by Y. The majorant is

$$\begin{aligned} \mathcal{M}_\oplus(v, y, \beta, \mu) := (1+\beta)(\mathcal{A}^{-1}y - \Lambda v, y - \mathcal{A}\Lambda v)_U + \\ + \left(1 + \frac{1}{\beta}\right) \frac{C_F^2}{c_1} \|\mu r\|^2_{\mathcal{V}} + (\mathcal{B}^{-1}(1-\mu)r, (1-\mu)r)_{\mathcal{V}}, \end{aligned} \tag{2.73}$$

where C_F and c_1 are from (2.38) and (2.36) respectively, and r is the residual,

$$r := f - \Lambda^* y - \mathcal{B}v.$$

The energy norm for the problem (2.34) is

$$|||\ w\ |||^2 := (\mathcal{A}\Lambda w, \Lambda w)_U + (\mathcal{B}w, w)_{\mathcal{V}}$$

and we recall

$$\|y\|^2_{\mathcal{A}} := (\mathcal{A}y, y)_U.$$

Theorem 2.4. *Let $v \in V_0$, and u be the solution of the problem* (2.34)*. Then*

$$|||\ u - v\ |||^2 \leq \mathcal{M}_\oplus(v, y, \beta, \mu), \quad \forall y \in Q,\ \beta > 0,\ \mu \in \mathrm{Y},$$

and

$$|||\ u - v\ |||^2 = \inf_{\substack{y \in Q, \\ \beta > 0}} \mathcal{M}_\oplus(v, y, \beta, 0).$$

Proof. Subtract $a(v, w)$ from the variational form and apply (2.40). Then

$$a(u - v, w) = (f - \mathcal{B}v, w)_{\mathcal{V}} - (\mathcal{A}\Lambda v, \Lambda w)_U - (\Lambda^* y, w)_{\mathcal{V}} + (y, \Lambda w)_U.$$

We reorganize

$$a(u - v, w) = (y - \mathcal{A}\Lambda v, \Lambda w)_U + (r, w)_{\mathcal{V}}.$$

We introduce the multiplier function μ which allows us to decompose

$$a(u-v,w) = (y - \mathcal{A}\Lambda v, \Lambda w)_U + (\mu r, w)_\mathcal{V} + ((1-\mu)r, w)_\mathcal{V}.$$

Now, we may estimate as in the proof of Theorem 2.3 and apply (2.45), Cauchy–Schwartz inequality, (2.36), and (2.38) to obtain

$$\begin{aligned}
a(u-v,w) &\leq (\mathcal{A}^{-1}y - \Lambda v, y - \mathcal{A}\Lambda v)_U^{\frac{1}{2}} \|\Lambda w\|_\mathcal{A} + \|\mu r\|_\mathcal{V} \frac{C_F}{\sqrt{c_1}} \|\Lambda w\|_\mathcal{A} + \\
&\quad + (\mathcal{B}^{-1}(1-\mu)r, (1-\mu)r)_\mathcal{V} (\mathcal{B}w, w)_\mathcal{V} \\
&= \left((\mathcal{A}^{-1}y - \Lambda v, y - \mathcal{A}\Lambda v)_U^{\frac{1}{2}} + \frac{C_F}{\sqrt{c_1}} \|\mu r\|_\mathcal{V} \right) \|\Lambda w\|_\mathcal{A} + \\
&\quad + (\mathcal{B}^{-1}(1-\mu)r, (1-\mu)r)_\mathcal{V}^{\frac{1}{2}} (\mathcal{B}w, w)_\mathcal{V}^{\frac{1}{2}} \\
&\leq \left(\left((\mathcal{A}^{-1}y - \Lambda v, y - \mathcal{A}\Lambda v)_U^{\frac{1}{2}} + \frac{C_F}{\sqrt{c_1}} \|\mu r\|_\mathcal{V} \right)^2 + (\mathcal{B}^{-1}(1-\mu)r, (1-\mu)r)_\mathcal{V} \right)^{\frac{1}{2}} ||| \, w \, ||| \,.
\end{aligned} \tag{2.74}$$

The rightmost term is the energy norm of w. Again we repeat the procedure of substituting $w := u - v$, squaring both sides and applying (2.10) to obtain

$$\begin{aligned}
||| \, u - v \, |||^2 &\leq \left((\mathcal{A}^{-1}y - \Lambda v, y - \mathcal{A}\Lambda v)_U^{\frac{1}{2}} + \frac{C_F}{\sqrt{c_1}} \|\mu r\|_\mathcal{V} \right)^2 + \\
&\quad + (\mathcal{B}^{-1}(1-\mu)r, (1-\mu)r)_\mathcal{V} \\
&\leq (1+\beta)(\mathcal{A}^{-1}y - \Lambda v, y - \mathcal{A}\Lambda v)_U + \\
&\quad + \left(1 + \frac{1}{\beta}\right) \frac{C_F^2}{c_1} \|\mu r\|_\mathcal{V}^2 + (\mathcal{B}^{-1}(1-\mu)r, (1-\mu)r)_\mathcal{V}.
\end{aligned}$$

□

Remarks similar to those discussed earlier for two types of reaction diffusion models are valid for the generalized problem (2.34). The function μ is at our disposal, if we select $\mu = 0$ or $\mu = 1$, we obtain following estimates of a special type:

$$||| \, u - v \, |||^2 \leq (\mathcal{A}^{-1}y - \Lambda v, y - \mathcal{A}\Lambda v)_U + (\mathcal{B}^{-1}r, r)_\mathcal{V} \tag{2.75}$$

and

$$||| \, u - v \, |||^2 \leq (1+\beta)(\mathcal{A}^{-1}y - \Lambda v, y - \mathcal{A}\Lambda v)_U + \left(1 + \frac{1}{\beta}\right) \frac{C_F^2}{c_1} \|r\|_\mathcal{V}^2. \tag{2.76}$$

The bound (2.75) has no "gap", i.e., for $y := p = \mathcal{A}\Lambda u$ it provides the exact deviation, whereas (2.76) does not have this property. The benefits of the bound (2.76) are evident if operator B^{-1} is such that it multiplies r significantly. Then, if y is not perfectly p, (2.75) may be very inaccurate, which is not the case with the bound (2.76).

If we select the optimal μ, the estimate obtained combines desired properties of (2.75) and (2.76). It is clear that the optimal μ is the minimizer of the quantity

$$I := \left(1 + \frac{1}{\beta}\right) \frac{C_F^2}{c_1} (\mu r, \mu r)_{\mathcal{V}}^2 + (\mathcal{B}^{-1}(1-\mu)r, (1-\mu)r)_{\mathcal{V}}.$$

Note that in the specific models of this type discussed earlier the computation of μ_{opt} can be done analytically.

2.4 Applications of deviation estimates

We summarize that deviation estimates (minorant and majorant) presented in Theorems 2.2, 2.3, and 2.4 are valid for any function $v \in V_0$, where V_0 is defined by the particular problem. The deviation estimates have a common form in which they depend explicitly on v, problem data, and some auxiliary functions (denoted here by w and y respectively). These estimates are natural tools for studying the "surroundings" of the solution in the energy space and provide a uniform basis for the evaluation of various types of errors.

- Consider the case where v is a solution of some simplified model and the minorant and majorant for the original model are at our disposal. Then they can be used to investigate the modelling error caused by the simplification of the model (see, e.g., [41, 42]).
- The explicit dependence of the deviation estimates on the data makes it possible to study the effects of incompletely known data, which is the main topic of the Thesis and is discussed in Chapter 4.
- In the case where v is some approximation of the exact solution u, these estimates are referred to as a posteriori error estimates. Regardless of the approximation method (no requirements for extra regularity or Galerkin orthogonality) by which v is obtained, these estimates can be used to investigate the approximation error.

In principle, it is possible to combine the error evaluations of different sources. Consider a case where we use some approximation method to compute an approximate solution of some simplified model for which the data are incompletely known. Using these deviation estimates it is possible to investigate quantitatively the errors generated by various sources. If we obtain this knowledge, then we can identify the main source of the error between the approximation we have and the exact solution of the original model. Moreover, we can balance the errors to avoid extensive computations, which are meaningless since the improved approximation accuracy is overshadowed by modelling and indeterminacy errors. If on top of this we have hierarchical models with gradually improved accuracy, we can balance the level of simplifications of the model applied, computational effort, and accuracy of the data.

A posteriori error estimates

The difference between the exact solution u of (2.35) and any approximation $v \in V_0$ can be estimated as follows:

$$\mathcal{M}_{\ominus}(v,w) \leq ||| \, u - v \, |||^2 \leq \mathcal{M}_{\oplus}(v,y), \quad \forall w \in V_0, \; y \in Q,$$

where w and y are at our disposal. Often, they are members of some finite dimensional subspaces, $w \in V_0^N \subset V_0$ and $y \in Q^N \subset Q$. The exact selection of basis functions generating V_0^N and Q^N depends on the problem type, computational resources, and the desired accuracy of estimates. For example, they can be piecewise polynomials with highly local support as in a traditional finite element approach.

In the computational practice the lower bound of the approximation error is of lesser interest than the upper bound. A simple way to compute an adequate value for the minorant is the following: Assume that we have computed an approximation on some finite basis; then we enrich the basis somehow and compute another approximation. Computing the difference of energies of these two approximations provides a lower bound for (one half of) the difference of the coarser approximation and the exact solution (in the energy norm).

For the upper bound, there are numerous variants of how to select the auxiliary function y. Recall that we would obtain the exact deviation if y is the exact stress $y = p$. In practice, p is not at our disposal. There are two principal ways to select the auxiliary function y:

(a) We postprocess the approximate solution (this procedure is denoted by $G : Q \to Q$) to obtain an approximation of the stress,

$$y := G(\mathcal{A}\Lambda v) \approx p.$$

(b) We minimize the majorant with respect to the auxiliary function y within some subspace $Q^N \in Q$, i.e., we solve the problem

$$\min_{\substack{y \in Q^N, \\ \beta > 0}} \mathcal{M}_{\oplus}(v, y, \beta).$$

The minimization procedure with respect to y and β can be done iteratively.

If it is necessary to obtain a reasonable upper bound with less computational effort, then method (a) is preferable. For a more accurate bound method (b) is recommended. Note that it not only provides an improved upper bound for the error, it also produces a good approximation of the true flux.

We present method (b) in more detail for the problem of type (2.35). The majorant is convex (quadratic) with respect to y. To compute the minimizer, it suffices to set the Gateaux derivative to zero. We recall the form of the majorant,

$$\mathcal{M}_{\oplus}(v,y,\beta) = \left(1 + \frac{1}{\beta}\right) \frac{C_F^2}{c_1} \|f - \Lambda^* y\|_{\mathcal{V}}^2 + \\ + (1+\beta)(\mathcal{A}^{-1}(\mathcal{A}y - \Lambda v), (y - \mathcal{A}\Lambda v))_U. \quad (2.77)$$

Two terms of the error majorant are related to the decomposed form of the classical equations. The first part is the error in the equilibrium condition (2.41) between an arbitrary function representing flux and the right hand side. We denote this part by

$$\mathcal{M}_{\oplus}^{\text{equi}} := \|f - \Lambda^* y\|^2_{\mathcal{V}}. \tag{2.78}$$

It guarantees the reliability of the estimate. The second part is the violation of the duality relation (2.42) between the approximation and arbitrary flux,

$$\mathcal{M}_{\oplus}^{\text{const}} := (\mathcal{A}^{-1}(y - \mathcal{A}\Lambda v), (y - \mathcal{A}\Lambda v))_U \tag{2.79}$$

If we substitute $y := p$ to the majorant, the second part provides the exact error and the first part is zero.

The necessary condition for the minimizer y can be computed as follows:

$$\mathcal{M}_{\oplus}(v, y + t\mu) = \left(1 + \frac{1}{\beta}\right)\frac{C_F^2}{c_1}\|f - \Lambda^* y - t\Lambda^*\mu\|^2_{\mathcal{V}} + \\ + (1+\beta)(\mathcal{A}^{-1}(y + t\mu - \mathcal{A}\Lambda v), (y + t\mu - \mathcal{A}\Lambda v))_U.$$

Therefore

$$\frac{\mathrm{d}\mathcal{M}_{\oplus}(v, y + t\mu)}{\mathrm{d}t} = \left(1 + \frac{1}{\beta}\right)\frac{C_F^2}{c_1}2((f - \Lambda^* y - t\Lambda^*\mu), -\Lambda^*\mu)_{\mathcal{V}} + \\ + (1+\beta)2(\mathcal{A}^{-1}(y + t\mu - \mathcal{A}\Lambda v), \mu)_U$$

and the condition

$$\left.\frac{\mathrm{d}\mathcal{M}_{\oplus}(v, y + t\mu)}{\mathrm{d}t}\right|_{t=0} = 0$$

reads

$$\left(1 + \frac{1}{\beta}\right)\frac{C_F^2}{c_1}(\Lambda^* y, \Lambda^*\mu)_{\mathcal{V}} + (1+\beta)(\mathcal{A}^{-1}y, \mu)_U = \\ \left(1 + \frac{1}{\beta}\right)\frac{C_F^2}{c_1}(f, \Lambda^*\mu)_{\mathcal{V}} + (1+\beta)(\mathcal{A}\Lambda v, \mu)_U. \tag{2.80}$$

Let y belong to a finite dimensional subspace of Q,

$$y \in \operatorname{span}\{\phi^1, \phi^2, \ldots, \phi^N\} =: Q_N \subset Q,$$

i.e.,

$$y = \sum_{i=1}^{N} \gamma_i \phi^i.$$

Then condition (2.80) leads to a system of linear equations,

$$\sum_{i=1}^{N} \gamma_i \left(\left(1 + \frac{1}{\beta}\right)\frac{C_F^2}{c_1}(\Lambda^*\phi^i, \Lambda^*\phi^j)_{\mathcal{V}} + (1+\beta)(\mathcal{A}^{-1}\phi^i, \phi^j)_U\right) = \\ \left(1 + \frac{1}{\beta}\right)\frac{C_F^2}{c_1}(f, \Lambda^*\phi^j)_{\mathcal{V}} + (1+\beta)(\mathcal{A}\Lambda v, \phi^j)_U, \quad j = \{1, \ldots N\}.$$

After introducing matrices

$$S_{ij}{}_{i,j=1}^{N} = (\Lambda^*\phi^i, \Lambda^*\phi^j)_{\mathcal{V}}, \qquad K_{ij}{}_{i,j=1}^{N} = (\mathcal{A}^{-1}\phi^i, \phi^j)_U,$$

and vectors

$$z_j{}_{j=1}^{N} = (f, \Lambda^*\phi^j)_{\mathcal{V}}, \qquad g_j{}_{j=1}^{N} = (\mathcal{A}\Lambda v, \phi^j)_U$$

the system can be written in the matrix form

$$\left(\left(1+\frac{1}{\beta}\right)\frac{C_F^2}{c_1}S + (1+\beta)K\right)\gamma = \left(1+\frac{1}{\beta}\right)\frac{C_F^2}{c_1}z + (1+\beta)g, \tag{2.81}$$

where γ is a vector consisting of the unknown coefficients. The evaluation of the majorant for $y \in Q_N$ can be easily done using the predefined matrices and the coefficient vector γ,

$$\mathcal{M}_{\oplus}(v,y,\beta) = \left(1+\frac{1}{\beta}\right)\frac{C_F^2}{c_1}\left(\gamma^T S\gamma - 2\gamma^T z + \|f\|_{\mathcal{V}}^2\right) + \\ + (1+\beta)\left(\gamma^T K\gamma - 2\gamma^T g + \|\Lambda v\|_{\mathcal{A}}^2\right).$$

If the majorant is minimized with respect to a positive scalar β, the minimum value is attained at

$$\beta := \left(\frac{\frac{C_F^2}{c_1}\|f-\Lambda^* y\|_{\mathcal{V}}^2\,\mathrm{ds}}{(\mathcal{A}^{-1}(y-\mathcal{A}\Lambda v),(y-\mathcal{A}\Lambda v))_U}\right)^{\frac{1}{2}}. \tag{2.82}$$

These observations motivate the Algorithm 1.

It is not obligatory to solve the equations (2.81) for y exactly. An efficient method for approximating y is the so called multi–grid method proposed in [54]. In general, iterative numerical methods for solving (2.81) are attractive alternatives, since at every iteration step one can compute the value of the majorant and cease all computations after the desired error estimation accuracy is obtained. It is rarely of interest to compute the value of the approximation error as accurately as possible; a reasonable upper bound for it is usually satisfactory. The construction and implementation of the error majorant has been studied for various problems in [2, 11, 12, 13, 14, 16].

For the computation of the majorant we need to estimate the constant C_F in inequality (2.38). In practice, to compute the majorant, we need only to roughly estimate the magnitude of the constant. For example, it can be estimated by solving approximately a generalized eigenvalue problem using the Galerkin approximation: Find eigenpairs (λ_i, v_i), where $v_i \in V_0^N \subset V_0$, such that

$$(\mathcal{A}\Lambda v_i, \Lambda w)_U = \lambda_i (v_i, w)_{\mathcal{V}}, \quad \forall w \in V_0^N. \tag{2.83}$$

The value of the constant is $C_F = \frac{1}{\lambda_{\min}}$, where $\lambda_{\min}$ is the lowest eigenvalue.

Algorithm 1 Minimization of the majorant.

Require: Matrices S and K, and vectors z and g are assembled and constants $||f||^2_{\mathcal{V}}$ and $\|\Lambda v\|^2_{\mathcal{A}}$ are computed. Set initial β_1.

for $k = 1$ to $I_{\max}$ **do**

Solve:

$$\left(\left(1+\frac{1}{\beta_k}\right)\frac{C_F^2}{c_1}S + (1+\beta_k)K\right)\gamma_{k+1} = \left(1+\frac{1}{\beta_k}\right)\frac{C_F^2}{c_1}z + (1+\beta_k)g.$$

Compute parts of the majorant:

$$\mathcal{M}_{\oplus}^{\text{equi}} = \gamma_{k+1}^T S\gamma_{k+1} - 2y_{k+1}^T z + \|f\|^2_{\mathcal{V}}$$
$$\mathcal{M}_{\oplus}^{\text{const}} = \gamma_{k+1}^T K\gamma_{k+1} - 2y_{k+1}^T g + \|\Lambda v\|^2_{\mathcal{A}}$$

Compute parameter β:

$$\beta_{k+1} = \left(\frac{\frac{C_F^2}{c_1}\mathcal{M}_{\oplus}^{\text{equi}}}{\mathcal{M}_{\oplus}^{\text{const}}}\right)^{\frac{1}{2}}$$

end for

If one wants to avoid solving the eigenvalue problem, one can use the following alternative method. We introduce a functional

$$G_\kappa(w) := \kappa(\mathcal{A}\Lambda w, \Lambda w)_U - \|w\|^2_{\mathcal{V}}.$$

It has the property

$$\inf_{w\in V_0} G_\kappa(w) = \left\{ \begin{array}{ll} -\infty, & \text{if } \kappa < C_F \\ \text{is positive}, & \text{if } \kappa \geq C_F \end{array} \right\}.$$

This functional can be used to obtain a rough estimate of C_F. The idea is to gradually increase κ until the minimum computed by some standard optimization technique (see, e.g., [4]) of $G_\kappa(w)$ over subspace $w \in V_0^N$ is positive.

Neither of the methodologies presented here to compute C_F is guaranteed. This observation is not related to any particular minimization method. In principle, no matter what κ we select, we can not guarantee that outside of our subspace V^N there is not a function for which G_κ attains a negative value. In practice, we can often have a reasonably good presentation of the space and obtain a decent approximation for the constant C_F. Moreover, it will be shown later with the paradigm of the Kirchhoff–Love arch problem that the minimized value of the majorant is not sensitive to even a large overestimation of C_F.

Besides the upper bound for the error, the majorant is the basis for various error indicators. It is relevant to distinguish between error indicators and estimators. The main goal of an error indicator is to form an adequate approximation of the error distribution with the lowest possible computational cost. This knowledge is essential for adaptive methods that iteratively enrich the set of basis functions used to compute the approximate solution. The theoretical basis dates back to [33]. In the last decades, error indicators have been intensively studied in numerical analysis and the amount of different methods and implementations is vast and beyond the scope of the Thesis (see, e.g., [1, 34, 50, 55, 58]).

3 ESTIMATES OF DEVIATION FOR KIRCHHOFF–LOVE ARCH

We consider a plane arch that has a constant cross section which is small compared to its length. Following [6], the arch and all related functions are presented in the parametrized form. The $\psi : [0,1] \to \mathbb{R}^2$ is a smooth parametrized non-self-intersecting curve of the curvilinear abscissa s that defines the shape of the arch. Displacement vector $u = (u_1, u_2)$ and load vector $f = (f_1, f_2)$ are given on a local basis (a_1, a_2) that varies along the arch, where a_1 is the tangential and a_2 is the normal direction. The angle between the horizontal axis and a_1 is denoted as θ. On both ends of the beam, there are known external loads, normal force N, shear force F, and bending moment M. The mentioned definitions with positive directions of the external loads are depicted in Figure 1. A more advanced formulation of the arch problem based on the control theory can be found in [31, 51], where regularity requirements for ψ are substantially relaxed.

The function $c : [0,1] \to \mathbb{R}$,

$$c(s) := \frac{1}{R(s)} = \frac{\psi_2''(s)\psi_1'(s) - \psi_1''(s)\psi_2'(s)}{(\psi_1'^2 + \psi_2^2)^{\frac{3}{2}}} \tag{3.1}$$

is the curvature of the arch. The energy functional of the arch is

$$J(u) = \frac{1}{2}\int_0^1 \left\{ EA(u_1' - cu_2)^2 + EI(cu_1 + u_2')'^2 \right\} \mathrm{ds}$$
$$- \int_0^1 f \cdot u \ \mathrm{ds} - \Big/_0^1 Nu_1 + \Big/_0^1 Fu_2 - \Big/_0^1 Mu_2', \tag{3.2}$$

where E is the material constant (Young's modulus), A is the area of the cross section and I is the second moment of inertia of the cross section. All these values are strictly positive. We apply the notation

$$\int_a^b f\,\mathrm{dx} = \Big/_a^b F = F(b) - F(a).$$

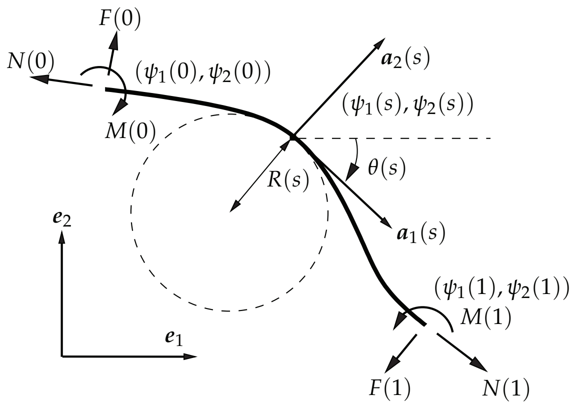

FIGURE 1 Kirchhoff-Love arch.

The minimizer of the energy functional is a solution $u \in V_0$ satisfying the integral relation

$$a(u,w) = l(w), \quad \forall\, w \in V_0, \tag{3.3}$$

where

$$a(u,w) = \int_0^1 \begin{bmatrix} EA(u_1' - cu_2) \\ EI(cu_1 + u_2')' \end{bmatrix} \cdot \begin{bmatrix} w_1' - cw_2 \\ (cw_1 + w_2')' \end{bmatrix} \,\mathrm{ds} \tag{3.4}$$

and

$$l(w) := \int_0^1 f \cdot w \;\,\mathrm{ds} + \Big/_0^1 Nw_1 - \Big/_0^1 Fw_2 + \Big/_0^1 Mw_2'. \tag{3.5}$$

If u is sufficiently regular, then (3.3) implies the classical equations

$$\begin{cases} -(EA(u_1' - cu_2))' - c(EI(cu_1 + u_2')')' &= f_1 \\ -cEA(u_1' - cu_2) + (EI(cu_1 + u_2')')'' &= f_2. \end{cases} \tag{3.6}$$

The boundary conditions are defined at the end points $s_1 = 0$ and $s_2 = 1$. They are listed as pairs in Table 2. Kinematic boundary conditions restrict displacement components or rotation and natural boundary conditions define tangential stress, shear force or bending moment. At both ends of the beam, either a natural or a corresponding kinematic boundary condition has to be defined in order to have a properly defined problem. Below, we are mainly concerned with cases where kinematic boundary conditions are homogeneous. Together with the regularity requirements, kinematic boundary conditions define the space of admissible displacements

$$V_0 := \{v \in V \mid v \text{ satisfies homogeneous kinematic boundary conditions}\}, \tag{3.7}$$

where we denote $V := H^1(0,1) \times H^2(0,1)$. The mentioned assumptions guarantee that in (3.5) either N,V, or M is known or the condition $w \in V_0$ implies the vanishing of the corresponding term.

TABLE 2 Boundary conditions of the Kirchhoff–Love arch.

kinematic		natural	
u_1	(tangential disp.)	N	(tangential stress)
u_2	(normal disp.)	F	(shear force)
u_2'	(rotation)	M	(bending moment)

For a straight beam (with $c = 0$), equations (3.6) are transformed to well known beam equations (see, e.g., [53]),

$$-EAu_1'' = f_1 \quad \text{and} \quad (EIu_2'')'' = f_2. \tag{3.8}$$

We note that (3.6) can be decomposed into

$$\left\{ \begin{array}{rcl} EA(u_1' - cu_2) & = & p_1 \\ EI(cu_1 + u_2')' & = & p_2 \end{array} \right. \tag{3.9}$$

and

$$\left\{ \begin{array}{rcl} -p_1' - cp_2' & = & f_1 \\ -cp_1 + p_2'' & = & f_2. \end{array} \right. \tag{3.10}$$

These relations are the physical laws governing the beam problem. Equations (3.9) define the constitutive relation that states the linear dependence between displacement u and tangential stress p_1 and the bending moment p_2. Henceforth, the vector p will be referred to as the stress vector. Equation (3.10) establishes the equilibrium between the external load f and the stresses p of the beam. Additionally, at the endpoints of the beam, stresses must satisfy the natural boundary conditions, namely

$$p_1 + cp_2 = N, \quad p_2 = M, \quad \text{and} \quad p_2' = F. \tag{3.11}$$

Stresses satisfying these relations form the space of admissible stresses,

$$Q_0 := \{y \in H^1(0,1) \times H^2(0,1) \mid y \text{ satisfies (3.11)}\}. \tag{3.12}$$

The problem is called statically determined (or overdetermined) if for $p = 0$, the equations (3.9) imply $u = 0$. It is not difficult to see that the kernel of equations (3.9) consists of rigid body motions (see [6]) that can be presented in the form

$$v = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} a\psi_2 + b_1 \\ a\psi_1 + b_2 \end{bmatrix}, \tag{3.13}$$

where a, b_1 and b_2 are constants. Thus, the beam is statically determined (or overdetermined) if the kinematic boundary conditions forbid any rigid body motion.

It is natural to classify boundary conditions into three main groups:

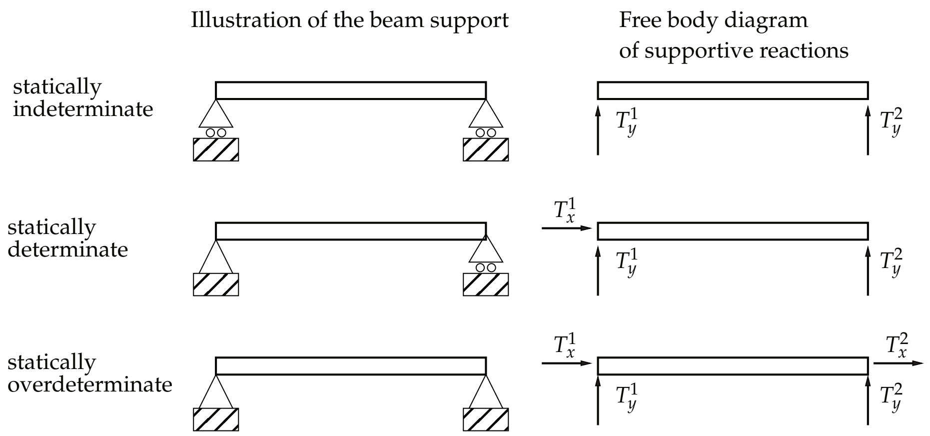

FIGURE 2 Examples of different boundary condition types.

- *Statically indetermined* cases, where there are not enough kinematic boundary conditions to restrict rigid body movement. The problem has no unique solution. These cases are considered non-physical and are neglected in our analysis.
- *Statically determined* cases, where there are three (dimension of the space of rigid body movements) kinematic boundary conditions that are sufficient to restrict rigid body movement. Additionally, no initial stresses are possible. In these cases the constitutive equations (3.9) and equilibrium equations (3.10) can be solved separately in a consequtive manner.
- *Statically overdetermined* cases, where there are more than three kinematic boundary conditions. In these cases equations (3.9) and (3.10) must be solved together as a single fourth order system. This kind of boundary conditions allows initial stresses for the unloaded beam. In other words, the null space of equation (3.10) has non-zero elements.

We demonstrate these types of boundary conditions by comparing the beams in Figure 2. Consider the beams in Figure 2 as rigid objects. For every beam, supportive reactions forbid any rotation on the plane and vertical movement, i.e., they do not move under moment or vertical force. In the statically indeterminate case, the supports allow rigid body movement, namely axial displacement. In the statically determinate and overdeterminate case this no longer happens, since the supports forbid axial movement. The physical difference between statically determined and overdetermined cases can be presented through an example of thermal elongation. Consider a beam that is initially at rest. Then the temperature increases and the beam elongates. In the statically determined case, the beam is allowed to elongate freely and no stresses appear. In the statically overdetermined case the beam experiences compression (and axial stress appears) due to the supports.

We can represent the Kirchhoff–Love arch model within the framework of the generalized variational problem introduced in Section 2.2. The constitutive relation (3.9) and the equilibrium relation (3.10) can be written as

$$\mathcal{A}\Lambda u = p \tag{3.14}$$

and

$$\Lambda^* p = f \tag{3.15}$$

respectively, where

$$\Lambda u := \begin{bmatrix} u_1' - cu_2 \\ (cu_1 + u_2')' \end{bmatrix}, \quad \Lambda^* p := \begin{bmatrix} -p_1' - cp_2' \\ -cp_1 + p_2'' \end{bmatrix}, \quad \text{and} \quad \mathcal{A} := \begin{bmatrix} EA & \\ & EI \end{bmatrix}. \tag{3.16}$$

With these definitions, the Kirchhoff–Love arch model has the form (2.35). For the existence of the solution we must show the ellipticity of $a : V_0 \times V_0 \to \mathbb{R}$ (see Section 2.2). For it, we need an analog of the inequality (2.30), which is proved in the book of Ciarlet [6] (Theorem 8.1.2., pg. 433),

Theorem 3.1. *If the function c is continuously differentiable over the interval I, the bilinear form*

$$a(u,v) = \int_I \{(u_1' - cu_2)(v_1' - cv_2) + (u_2' + cu_1)'(v_2 + cu_1)'\} \text{ ds}$$

is $H_0^1(I) \times (H^2(I) \cap H_0^1(I))$-elliptic, and thus, it is a fortiori $H_0^1(I) \times H_0^2(I)$-elliptic.

In other words, Theorem 3.1 states that for a statically determinate or overdeterminate beam, there exists a positive constant C such that

$$\int_0^1 \left\{ w_1^2 + w_2^2 + w_1'^2 + w_2'^2 + w_2''^2 \right\} \text{ ds} \leq C \int_0^1 \left\{ (w_1' - cw_2)^2 + (cw_1 + w_2')'^2 \right\} \text{ ds}, \tag{3.17}$$

for all $w \in V_0$.

3.1 Estimates of deviations for the Kirchhoff–Love arch model

The estimates of deviations for the Kirchhoff–Love arch model were first presented in [24]. They have a similar structure as the general estimates presented in Section 2.3. The majorant can be derived with the help of integral identities (as has been done in Theorem 2.3 in the context of a general variation problem of type (2.34)).

Theorem 3.2. *Let u be a solution of* (3.3) *and $v \in V_0$. Then*

$$|\!|\!| u - v |\!|\!|^2 \leq \mathcal{M}_\oplus(v, y, \beta), \quad y \in Q_0, \ \beta > 0,$$

where

$$\mathcal{M}_{\oplus}(v,y,\beta) := \left(1+\frac{1}{\beta}\right)\frac{C}{\alpha}\int_0^1 \left\{(f_1+(y_1'+cy_2'))^2+(f_2-(cy_1+y_2''))^2\right\} \mathrm{ds}+$$

$$+(1+\beta)\int_0^1 \left\{\frac{1}{EA}(y_1-EA(v_1'-cv_2))^2+\frac{1}{EI}(y_2-EI(cv_1+v_2')'^2\right\} \mathrm{ds}, \quad (3.18)$$

where C is from (3.17), $\alpha := \min\{EA, EI\}$, *and*

$$\| w \|^2 := \frac{1}{2}\int_0^1 \left\{EA(w_1'-cw_2)^2+EI(cw_1+w_2')'^2\right\} \mathrm{ds}.$$

Proof. We begin with the integral identity that defines the generalized solution of the arch problem. We note that

$$\int_0^1 \begin{bmatrix} w_1'-cw_2 \\ (cw_1+w_2')' \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \mathrm{ds} = \int_0^1 \begin{bmatrix} -y_1'-cy_2' \\ -cy_1+y_2'' \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \mathrm{ds}+$$

$$+ \Big/_0^1 w_1y_1 + \Big/_0^1 (cw_1+w_2')y_2 - \Big/_0^1 w_2y_2' \quad (3.19)$$

for any $w \in H^1(0,1)\times H^2(0,1)$ and $y \in H^1(0,1)\times H^2(0,1)$.

By (3.3) and (3.19) we obtain

$$a(u-v,w) = \int_0^1 f\cdot w\,\mathrm{ds} + \Big/_0^1 Nw_1 - \Big/_0^1 Fw_2 + \Big/_0^1 Mw_2' -$$

$$-\int_0^1 \begin{bmatrix} EA(v_1'-cv_2) \\ EI(cv_1+v_2')' \end{bmatrix} \cdot \begin{bmatrix} w_1'-cw_2 \\ (cw_1+w_2')' \end{bmatrix} \mathrm{ds} + \int_0^1 \begin{bmatrix} w_1'-cw_2 \\ (cw_1+w_2')' \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \mathrm{ds} -$$

$$-\int_0^1 \begin{bmatrix} -y_1'-cy_2' \\ cy_1+y_2'' \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \mathrm{ds} - \Big/_0^1 w_1y_1 - \Big/_0^1 (cw_1+w_2')y_2 + \Big/_0^1 w_2y_2'. \quad (3.20)$$

We rewrite (3.20) in the form

$$a(u-v,w) = I_1+I_2+I_3, \quad (3.21)$$

where

$$I_1 = \int_0^1 \begin{bmatrix} f_1+(y_1'+cy_2') \\ f_2-(cy_1+y_2'') \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \mathrm{ds},$$

$$I_2 = \int_0^1 \begin{bmatrix} y_1-EA(v_1'-cv_2) \\ y_2-EI(cv_1+v_2')' \end{bmatrix} \cdot \begin{bmatrix} w_1'-cw_2 \\ (cw_1+w_2')' \end{bmatrix} \mathrm{ds},$$

and

$$I_3 = \int_0^1 (N - y_1 - cy_2)w_1 + \int_0^1 (-F + y_2')w_2 + \int_0^1 (M - y_2)w_2'.$$

After imposing boundary conditions, $w \in V_0$ and $y \in Q_0$, I_3 vanishes.

By the Cauchy-Schwartz inequality, we have

$$I_1 \quad \leq \quad \left(\int_0^1 (f_1 + (y_1' + cy_2'))^2 + (f_2 - (cy_1 + y_2''))^2 \ \mathrm{ds} \right)^{\frac{1}{2}} \left(\int_0^1 w_1^2 + w_2^2 \ \mathrm{ds} \right)^{\frac{1}{2}}. \tag{3.22}$$

We can estimate the L_2-norm of w from above by the Sobolev norm of $H^1(0,1) \times H^2(0,1)$ and apply (3.17), then

$$I_1 \leq \left(\int_0^1 (f_1 + (y_1' + cy_2'))^2 + (f_2 - (cy_1 + y_2''))^2 \ \mathrm{ds} \right)^{\frac{1}{2}} \times$$
$$\times \frac{\sqrt{C}}{\sqrt{\alpha}} \left(\int_0^1 EA(w_1' - cw_2)^2 + EI(cw_1 + w_2')'^2 \ \mathrm{ds} \right)^{\frac{1}{2}}, \tag{3.23}$$

where $\alpha = \min\{EA, EI\}$. Now, we apply the Cauchy-Schwartz inequality again,

$$I_2 = \int_0^1 \begin{bmatrix} \frac{1}{\sqrt{EA}}(y_1 - EA(v_1' - cv_2)) \\ \frac{1}{\sqrt{EI}}(y_2 - EI(cv_1 + v_2')') \end{bmatrix} \cdot \begin{bmatrix} \sqrt{EA}(w_1' - cw_2) \\ \sqrt{EI}(cw_1 + w_2')' \end{bmatrix} \mathrm{ds}$$
$$\leq \left(\int_0^1 \frac{1}{EA}(y_1 - EA(v_1' - cv_2))^2 + \frac{1}{EI}(y_2 - EI(cv_1 + v_2')'^2 \mathrm{ds} \right)^{\frac{1}{2}} \parallel\!| \ w \ |\!\parallel. \tag{3.24}$$

These estimates have $|\!|\!| \ w \ |\!|\!|$ as a common multiplier. Apply them to (3.21) and set $w = u - v$. Then on the left hand side we obtain $|\!|\!| \ u - v \ |\!|\!|^2$ and we may divide by $|\!|\!| \ u - v \ |\!|\!|$ on both sides. Then we arrive at

$$|\!|\!| \ u - v \ |\!|\!| \leq \frac{\sqrt{C}}{\sqrt{\alpha}} \left(\int_0^1 (f_1 + (y_1' + cy_2'))^2 + (f_2 - (cy_1 + y_2''))^2 \ \mathrm{ds} \right)^{\frac{1}{2}} +$$
$$+ \left(\int_0^1 \frac{1}{EA}(y_1 - EA(v_1' - cv_2))^2 + \frac{1}{EI}(y_2 - EI(cv_1 + v_2')'^2 \mathrm{ds} \right)^{\frac{1}{2}}. \tag{3.25}$$

For computation purposes it is preferable to have quadratic expressions. Thus we introduce arbitrary $\beta \in \mathbb{R}_+$ and use Young's inequality (2.10) to obtain

$$\| u - v \|^2 \leq \left(1 + \frac{1}{\beta}\right) \frac{C}{\alpha} \int_0^1 \left\{ (f_1 + (y_1' + cy_2'))^2 + (f_2 - (cy_1 + y_2''))^2 \right\} \mathrm{ds} +$$

$$+ (1+\beta) \int_0^1 \left\{ \frac{1}{EA}(y_1 - EA(v_1' - cv_2))^2 + \frac{1}{EI}(y_2 - EI(cv_1 + v_2')'^2 \right\} \mathrm{ds}. \quad (3.26)$$

The right hand side will be called the error majorant (or majorant) and will be denoted by $\mathcal{M}_\oplus(v, y, \beta)$. □

Remark 3.1. *Two terms of the error majorant are related to the decomposed form of the classical equations. The first part is the error in the equilibrium condition* (3.10) *between an arbitrary function representing stresses and the external load. The second part is the violation of the duality relation* (3.9) *between the approximation and arbitrary stresses. If we substitute exact stresses $y := p$ to the majorant, the second part provides the exact error and the first part is zero.*

Remark 3.2. *Under the definitions* (3.16) *the majorant has the typical structure*

$$\mathcal{M}_\oplus(v, y, \beta) := \left(1 + \frac{1}{\beta}\right) \frac{C}{\alpha} \int_0^1 |f - \Lambda^* y|^2 \,\mathrm{ds} +$$

$$+ (1+\beta) \int_0^1 \mathcal{A}^{-1}(y - \mathcal{A}\Lambda v) \cdot (y - \mathcal{A}\Lambda v) \,\mathrm{ds}. \quad (3.27)$$

Theorem 3.3. *Let u be a solution of* (3.3) *and $v \in V_0$. Then*

$$\| u - v \|^2 \geq \mathcal{M}_\ominus(v, w), \quad \forall w \in V_0,$$

where

$$\mathcal{M}_\ominus(v, w) := -\int_0^1 \begin{bmatrix} EA(w_1' - cw_2 + 2(v_1' - cv_2)) \\ EI((w_2' + cw_1')' + 2(v_2' + cv_1)') \end{bmatrix} \cdot \begin{bmatrix} w_1' - cw_2 \\ (w_2' + cw_1')' \end{bmatrix} \mathrm{ds} +$$

$$+ 2 \int_0^1 (f_1 w_1 + f_2 w_2) \;\mathrm{ds}. \quad (3.28)$$

Proof. We note that by applying definitions (3.16) the energy functional of the arch problem can be written as in (2.28). By Theorem 2.2, we obtain the minorant of the following form:

$$\mathcal{M}_\ominus(v, w) := -(\mathcal{A}\Lambda w, \Lambda(w + 2v))_U + 2(f, w)_\mathcal{V}.$$

Applying the definitions (3.16), we can write the minorant explicitly for the Kirchhoff–Love arch problem. □

TABLE 3 Approximations of the constant C in (3.17) by solving the generalized eigenvalue problem (2.83) using the Galerkin method, where the subspace V_0^N is defined in (3.29).

N	2	4	6	8	10
C	0.050661	0.050661	0.050661	0.050661	0.050661

3.2 Numerical example: uniformly curved beam

The Kirchhoff–Love arch problem is of the general type (2.35). We recall the discussion in Section 2.4 and apply different methods presented there to estimate the approximation error.

We consider a half circular beam

$$\psi(t) = \begin{bmatrix} \cos(\pi t) \\ \sin(\pi t) \end{bmatrix}, \quad t \in [0,1],$$

where the curvature is $c = 1$. Let both ends of the beam be clamped, i.e., the displacement satisfies the boundary conditions

$$u_1(0) = u_2(0) = u_2'(0) = u_1(1) = u_2(1) = u_2'(1) = 0.$$

We normalize $EA = EI = 1$.

First, we compute an approximation of the constant C in (3.17). The basis that satisfies the boundary conditions can be easily constructed using Fourier type basis functions. Let

$$w \in V_0^N := \operatorname{span} \left\{ \begin{bmatrix} \sin(k\pi t) \\ 0 \end{bmatrix}, \begin{bmatrix} 1 - \cos(2k\pi t) \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 - \cos(2k\pi t) \end{bmatrix} \right\}_{k=1}^{N}. \tag{3.29}$$

Then, $\dim(V_0^N) = 3N$. We approximate C by solving the general eigenvalue problem (2.83) using the Galerkin method. In Table 3, we increased the number of basis functions and observe how the approximated value for C develops. Results indicate that the eigenmode corresponding to the lowest eigenvalue is within our basis. We estimate from above and set $C = 1$. We will show later (in Table 5) that even a very crude overestimation of the constant C has no significant effect on the computed value of the majorant.

A posteriori error estimates

We introduce a polynomial solution

$$u(t) = \begin{bmatrix} t(t-1) \\ t^2(t-1)^2 \end{bmatrix}, \tag{3.30}$$

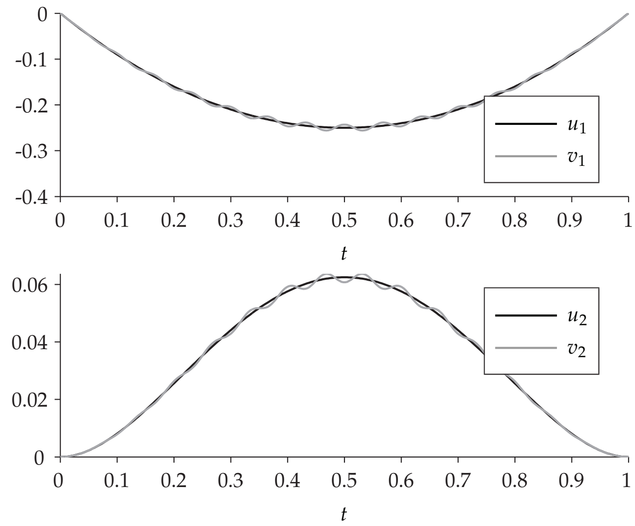

FIGURE 3 Exact solution u and "approximation" v.

which satisfies the kinematic boundary conditions. From (3.9) and (3.10), we can easily compute stress

$$p(t) = \begin{bmatrix} EA\left(-t^4+2t^3-t^2+2t-1\right) \\ EI\left(12t^2-10t+1\right) \end{bmatrix}$$

and load

$$f(t) = \begin{bmatrix} EA\left(4t^3-6t^2+2t-2\right)+EI(-24t+10) \\ EA\left(t^4-2t^3+t^2-2t+1\right)+24EI \end{bmatrix}.$$

To study the application of a posteriori error estimates to control the approximation error, we define an "approximate solution" v of the form

$$v := u + \epsilon\xi,$$

where ξ is a known "error" that satisfies the kinematic boundary conditions. We selected

$$\xi := \begin{bmatrix} t(t-1)\cos(30\pi t) \\ t^2(t-1)^2\cos(30\pi t) \end{bmatrix}.$$

We set $\epsilon := 0.02\frac{\|u\|}{\|\xi\|}$ to obtain the following relative error (in the L_2-norm):

$$\frac{\|u-v\|}{\|u\|} = \epsilon\frac{\|\xi\|}{\|u\|} = 2\%.$$

The exact solution and the "approximation" are depicted in Figure 3.

To measure the efficiency of the a posteriori error estimates, we introduce efficiency indexes,

$$I_{\text{eff}}^{\oplus} := \frac{\mathcal{M}_{\oplus}}{\|\!| u-v |\!\|^2}. \tag{3.31}$$

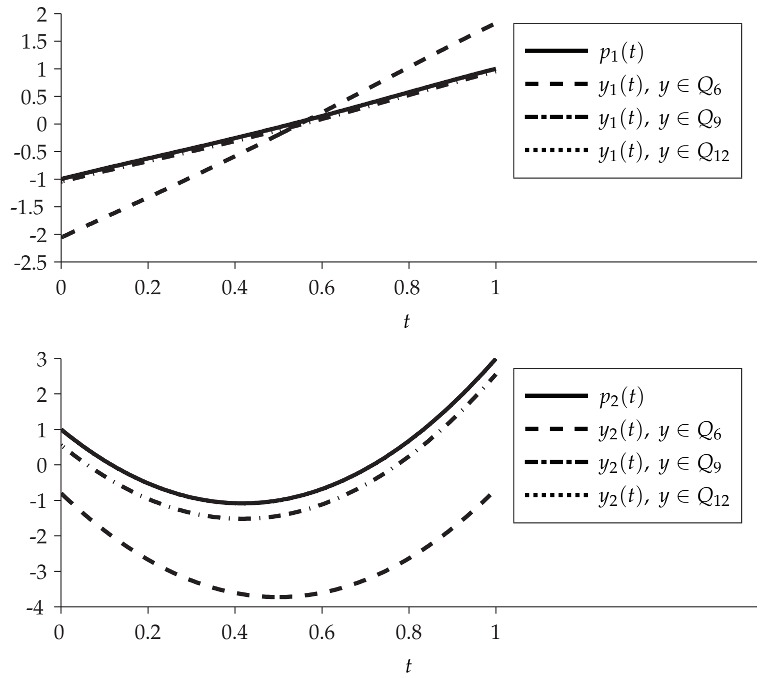

FIGURE 4 The approximate y obtained through Algorithm 1, where the dimension of y varies, compared to the exact stress p.

and

$$I_{\text{eff}}^{\ominus} := \frac{\mathcal{M}_{\ominus}}{|||\ u - v\ |||^2}. \tag{3.32}$$

Since deviation estimates are guaranteed,

$$I_{\text{eff}}^{\ominus} \leq 1 \leq I_{\text{eff}}^{\oplus}$$

For our test case we have $|||\ u - v\ |||^2 = 50.740$.

Let y in the majorant be defined on a Fourier basis, i.e.,

$$y \in Q_N := \text{span}\left\{ \begin{bmatrix} \sin(k\pi t) \\ 0 \end{bmatrix}, \begin{bmatrix} \cos(k\pi t) \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \sin(k\pi t) \end{bmatrix}, \begin{bmatrix} 0 \\ \cos(k\pi t) \end{bmatrix} \right\}_{k=1}^{N}.$$

Note that $\dim(Q_N) = 4N$. We minimize the majorant with respect to $y \in Q_N$ following Algorithm 1. Regardless of the dimension N the iteration converged within five to six steps. In Figure 4, we have depicted y and the exact stress p. Clearly, y approaches p as the dimension of the space Q_N increases. For $N = 12$ the difference between the curves is no longer visible. In Table 4, we can observe how the majorant improves as N increases. The efficiency index tends to one as the majorant approaches the exact deviation. Moreover, the equilibrium part of the majorant tends to zero as the constitutive part approaches the exact deviation error.

TABLE 4 Efficiency index of the majorant with different values of N.

N	4	6	8	9	10	11	12
$I_{\text{eff}}^{\oplus}$	1.1853	1.1509	1.0204	1.0048	1.0011	1.0002	1. 0001
$\mathcal{M}_{\oplus}$	60.143	58.398	51.774	50.985	50.7949	50.752	50.742
$\mathcal{M}_{\oplus}^{\text{equi}}$	0.01320	0.00229	0.00485	0.00029	0.00002	$7.2 \cdot 10^{-7}$	$3.3 \cdot 10^{-8}$
$\mathcal{M}_{\oplus}^{\text{const}}$	58.374	57.658	50.776	50.742	50.740	50.740	50.740

TABLE 5 Efficiency index of the majorant with different values of N and the constant C.

N		4	6	8	9	10	11	12
$C = 1$	$I_{\text{eff}}^{\oplus}$	1.1853	1.1509	1.0204	1.0048	1.0011	1.0002	1. 0001
$C = 10$	$I_{\text{eff}}^{\oplus}$	1.2587	1.1601	1.0601	1.0151	1.0034	1.0008	1.0002
$C = 100$	$I_{\text{eff}}^{\oplus}$	1.4992	1.1747	1.1413	1.0453	1.0108	1.0024	1.0005
$C = 1000$	$I_{\text{eff}}^{\oplus}$	2.3947	1.2026	1.1592	1.1187	1.0329	1.0075	1.0016

TABLE 6 Efficiency index of the minorant with different values of N.

N	2	3	4	5	6
$I_{\text{eff}}^{\ominus}$	0.99971	0.99989	0.99995	0.99997	0.99998
$\mathcal{M}_{\ominus}(v, w^N)$	50.725	50.7341	50.737	50.7382	50.7388

Remark 3.3. *Since the term related to the equilibrium relation tends to zero, even a substantial overestimation of the constant C does not seriously affect the efficiency of these estimates. In Table 5, we show the efficiency indexes obtained by different values for the constant C.*

Next, we study the minorant. We solve problem (3.3) using the Galerkin method with the Fourier type subspace V_0^N in (3.29). Then, we compute the energy of the obtained approximation w^N and estimate the error from below by comparing it to the energy of v as follows:

$$\| u - v \|^2 \geq 2 \left(J(v) - J(w^N) \right) =: \mathcal{M}_{\ominus}(v, w^N). \tag{3.33}$$

The resulting lower bounds are presented in Table 6. The reason why it is so efficient is that applied basis (3.29) can represent the exact solution (3.30) very well with a small number of basis functions. Note that the inequality in (3.33) becomes an equality if $w := u$.

Alternatively, a lower bound can be computed directly by maximizing the minorant in Theorem 2.2. Recall that it provides the exact error if $w := u - v$. For our model problem and for applied basis (3.29), the exact error is much more complicated to approximate, since it contains high frequency oscillations. This obstacle can be overcome if one uses locally supported basis functions.

4 INDETERMINACY OF THE DATA

In this Chapter we consider elliptic boundary value problems defined in Section 2.1, where the data are incompletely known. Motivated by engineering practice, we assume that the data are given in the form of mean value and variations. For example, the heat conduction coefficient can be given in the form

$$a = a_\circ \pm \delta.$$

The accuracy $\left|\frac{\delta}{a_\circ}\right|$ depends on the measurement methods, the homogeneity of the material, etc. Similar uncertainty in the coefficients arises in many other problems, for example linear elasticity and magnetostatics.

Henceforth, the set of admissible data is denoted by $\mathcal{D}$. The known "mean" data are denoted by $D_\circ$. The subscript $\circ$ denotes the data (and solution) associated to the mean data. The mean value together with allowed perturbations will define a set of admissible data. The set of admissible data (i.e., the mean value denoted by $D_\circ$ and the data within given variations) will be denoted by $\mathcal{D}$. We assume $\mathcal{D}$ to be such that the problems with any data from $\mathcal{D}$ are well defined, i.e., they possess unique solution. Moreover, we define a solution mapping

$$\mathcal{S} : \mathcal{D} \to \mathcal{S}(\mathcal{D}) \subset V_0,$$

which defines the solution resulting from particular data. Note that the space of admissible solutions is fixed. In other words, with all admissible data the solution is in V_0. Thus, we will for example not allow variations on the subset of boundary where Dirichlet boundary conditions are set.

The solution set $\mathcal{S}(\mathcal{D})$ is of particular interest. It is the set of admissible solutions, i.e., solutions that may result under the existing information under the possible data. Our analysis investigates two related topics:

- **Relations between sets $\mathcal{D}$ and $\mathcal{S}(\mathcal{D})$.** To measure the set $\mathcal{S}(\mathcal{D})$, we choose $u_\circ = \mathcal{S}(D_\circ)$ as an anchor and use it to define the *radius of the solution set* as follows:

$$r := \sup_{u \in \mathcal{S}(\mathcal{D})} ||| \, u_\circ - u \, |||_\circ \tag{4.1}$$

and its normalized counterpart,

$$\bar{r} := \sup_{u \in \mathcal{S}(\mathcal{D})} \frac{\| u_\circ - u \|_\circ}{\| u_\circ \|_\circ}, \tag{4.2}$$

where $\| \cdot \|_\circ$ denotes the energy norm generated by the data $D_\circ$. Of course the radius can be defined by a dual variable, energy, or by any functional of u.

- **The sensitivity of the solutions and related quantities with respect to variations in data.** The goal of the study is the derivation of various quantities of interest with respect to the data values. It is of special interest to see how this derivative depends on the value of $D_\circ$, and what other factors characterize its magnitude.

Additionally, we define the error quantities related to some particular $v \in V_0$. The distance of v from the mean solution is denoted by

$$e_\circ := \| u_\circ - v \|_\circ .$$

Since the exact solution is not uniquely defined, it is natural to define the worst case error

$$e_{\text{worst}} := \sup_{u \in \mathcal{S}(D)} \| u - v \|_\circ \tag{4.3}$$

and the best case error

$$e_{\text{best}} := \inf_{u \in \mathcal{S}(D)} \| u - v \|_\circ . \tag{4.4}$$

In the case of incompletely known data, the standard definition of the approximation error is not applicable. If the distance from v to set $\mathcal{S}(\mathcal{D})$ is much larger than $\operatorname{diam}(\mathcal{S}(\mathcal{D}))$, (then $e_\circ \approx e_{\text{worst}} \approx e_{\text{best}}$) then the distance from v to any member of the solution set represents well the approximation error. If this is not the case, i.e., $e_{\text{worst}} \gg e_{\text{best}}$ then any further efforts to improve the approximative solution v are useless. The definitions are illustrated in Figures 5 and 6.

4.1 Effects of indeterminate data in algebraic problems

Below, we consider several elementary problems which allow us to demonstrate effects arising from an incomplete knowledge of the data. These effects are of different types. In real life mathematical models, they may arise in more complicated forms.

4.1.1 Example 1

Consider a two-dimensional minimization problem,

$$\min_{x \in \mathbb{R}^2} \left\{ \tfrac{1}{2} x^T A x - b^T x \right\}, \tag{4.5}$$

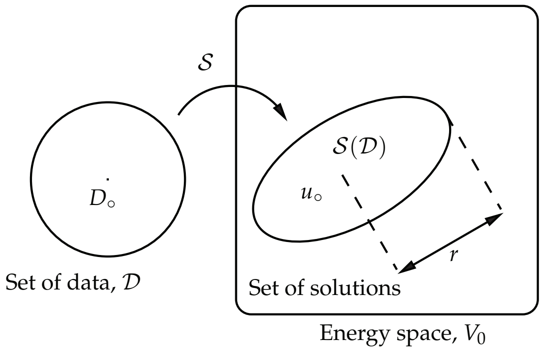

FIGURE 5 Illustration of the radius of the solution set.

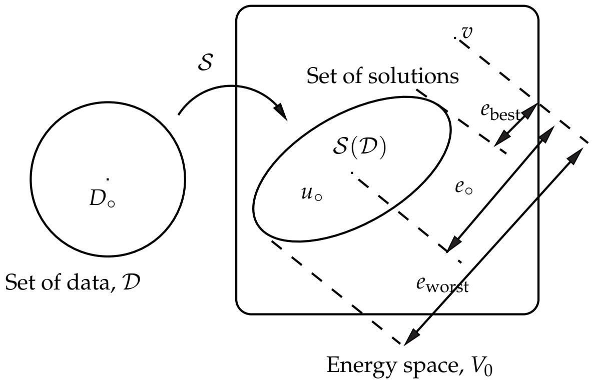

FIGURE 6 Illustration of different distances associated to $v \in V_0$.

where $A \in \mathbb{R}^{2\times 2}$ is positive definitive and symmetric, and $b \in \mathbb{R}^2$, which is equivalent to the system

$$Ax = b.$$

Let us now assume that the matrix A is not fully determined, but we know that it belongs to the set

$$\mathcal{D}_A := \{A \in \mathbb{M}^{2\times 2} \,|\, A = A_\circ + \delta E,\ E \in \mathbb{M}^{2\times 2},\ |E| \le 1,\ \delta > 0\}$$

of "possible" matrices. Here, δ is the magnitude of the perturbation, which can be considered as a measure of the indeterminacy. We denote the condition number of $A_\circ$ by

$$\text{cond}(A_\circ) := \frac{\overline{c}}{\underline{c}},$$

where $\underline{c}$ and $\overline{c}$ are the minimal and maximal eigenvalues, respectively, so that

$$\underline{c}|x|^2 \le A_\circ x \cdot x \le \overline{c}|x|^2, \quad \forall x \in \mathbb{R}^2.$$

We assume that

$$\delta \le \underline{c}, \tag{4.6}$$

which guarantees that all the matrices in $\mathcal{D}_A$ are positive definite. The absolute value of the perturbation magnitude δ alone can not tell how significant the range of indeterminacy is. For this reason, we introduce the parameter

$$\theta = \frac{\delta}{\underline{c}}, \tag{4.7}$$

which characterizes a "normalized" perturbation and serves as a more adequate measure of indeterminacy[1].

It is easy to show that the larger the condition number of a matrix is, the more sensitive (with respect to small perturbations) the problem (4.5) is. For example, consider matrices

$$A_\circ^{(1)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad A_\circ^{(2)} = \frac{1}{3}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

The matrix $A_\circ^{(1)}$ has $\underline{c} = \overline{c} = 1$ and $A_\circ^{(2)}$ has $\underline{c} = \frac{1}{3}$ and $\overline{c} = 1$. Let $\delta = 0.05$. For this algebraic problem, it is easy to explicitly compute the solution set

$$\mathcal{S}(\mathcal{D}_A) := \{x = A^{-1}b \mid A \in \mathcal{D}_A\} \tag{4.8}$$

and define the radius

$$r := \sup_{x \in \mathcal{S}(\mathcal{D}_A)} \frac{|x_0 - x|}{|x_0|}.$$

We chose $b := [1,1]^T$, so that the non–perturbed solution is $x_0 = [1,1]^T$ for both $A_\circ^{(1)}$ and $A_\circ^{(2)}$. Figure 7 depicts the solutions for systems

$$\left(A_\circ^{(1)} + \delta E\right) x = b \quad \text{and} \left(A_\circ^{(2)} + \delta E\right) x = b$$

for 10 000 randomly computed E. It is easy to see that the set $\mathcal{S}(\mathcal{D}_A)$ is much larger for the matrix $A_\circ^{(2)}$: $r^{(1)} = 0.053$ and $r^{(2)} = 0.152$. The condition number 3 of $A^{(2)}$ is actually rather small. We note that in real applications (e.g., for matrices generated by strongly non–isotropic diffusion), these effects may become much more significant.

4.1.2 Example 2

Let the matrix A contain a parameter α, which is known to be inside some interval, e.g.,

$$A_\alpha = \begin{bmatrix} 2-\alpha & 1 \\ 1 & 2-\alpha \end{bmatrix}, \quad \alpha \in [\alpha_\circ - \delta, \alpha_\circ + \delta] \subset (1,2],$$

[1] We note that a normalization by the matrix norm is not proper, since we want to obtain a quantity that indicates how serious the perturbation δ can be. Clearly, it is most critical when $A_\circ$ is "smallest", i.e., when it multiplies vectors close to the eigenvector of the smallest eigenvalue $\underline{c}$.

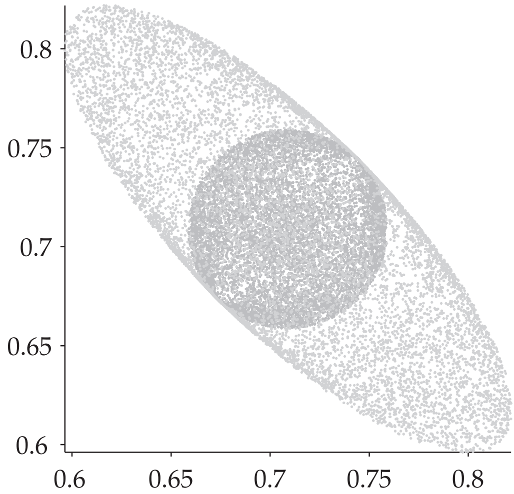

FIGURE 7 Solution sets (4.8) for matrices $A_\circ^{(1)}$ (darker) and $A_\circ^{(2)}$ (lighter), 10 000 random perturbations E were applied to illustrate the solution set.

where δ is the accuracy within which the mean value $\alpha_\circ$ is known. By x_α we denote the solution of the system,

$$A_\alpha x_\alpha = f.$$

Since $\det(A_\alpha) = (2-\alpha)^2 - 1$, the matrix is non-degenerate for any $\alpha \neq 1$ (and $\alpha \neq 3$), but if $\alpha = 1$ (or $\alpha = 3$), then the matrix has a nontrivial kernel. The behavior of the solution set as $\alpha_\circ \to 1$ is of our interest. The inverse matrix to A_α is

$$A_\alpha^{-1} = \frac{1}{(2-\alpha)^2 - 1} \begin{bmatrix} 2-\alpha & -1 \\ -1 & 2-\alpha \end{bmatrix}.$$

Hence,

$$x_\alpha = A_\alpha^{-1} f = \frac{1}{(2-\alpha)^2 - 1}((1-\alpha)I + T)f = \frac{1}{3-\alpha} f + \frac{1}{(\alpha-1)(\alpha-3)} Tf,$$

where

$$T = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

is a matrix with non-zero kernel $\mathcal{N}(T)$. Since

$$Tf = \begin{bmatrix} f_1 - f_2 \\ -f_1 + f_2 \end{bmatrix},$$

Tf is a zero vector if and only if $f_1 = f_2$, which is equivalent to $f \in \mathcal{N}(T)$. This condition is crucial for the stability of the solution with respect to small changes of α.

TABLE 7 Limits of the derivative and logarithmic derivatives of x_α.

	$\lim_{\alpha\to 1} \left\lvert \frac{\partial x_\alpha}{\partial \alpha} \right\rvert$	$\lim_{\alpha\to 1} \frac{1}{\lvert x_\alpha \rvert} \left\lvert \frac{\partial x_\alpha}{\partial \alpha} \right\rvert$
$f \notin \mathcal{N}(T)$	∞	∞
$f \in \mathcal{N}(T)$	$\frac{1}{4}\lvert f \rvert$	$\frac{1}{2}$

Let us study the behavior of the solution with respect to small changes of the unknown parameters. For this purpose we define the increment

$$\Delta_\delta^\alpha x_\alpha := \frac{x_{\alpha+\delta} - x_\alpha}{\delta} \tag{4.9}$$

and the normalized increment $\frac{1}{|x_\alpha|}\Delta_\delta^\alpha x_\alpha$. It is clear that if $\frac{1}{|x_\alpha|}\Delta_\delta^\alpha x_\alpha$ is very large with small values of δ, then the solutions generated by the different α are significantly different. If $\delta \to 0$, these quantities tend to the derivative $\frac{\partial x^\alpha}{\partial \alpha}$ and the logarithmic derivative $\frac{1}{|x^\alpha|}\frac{\partial x^\alpha}{\partial \alpha}$, respectively. We note that the logarithmic derivative is a fundamental quantity in the sensitivity analysis.

We have

$$\frac{\partial x_\alpha}{\partial \alpha} = \frac{1}{(3-\alpha)^2} f + \frac{4-2\alpha}{(\alpha-1)^2(\alpha-3)^2} Tf.$$

If $Tf = 0$, then the second term vanishes and $\frac{\partial x_\alpha}{\partial \alpha}$ is relatively insensitive with respect to variations of α close to one. However, if $Tf \neq 0$, then the second term is very sensitive if α is close to one. These observations are summarized in Table 7.

For practical computations this qualitative observation is important, but not sufficient. For a reliable quantitative analysis, two other questions are of utmost importance:

(a) For a given indeterminacy in the data, how large is the inaccuracy in the solution (or other quantity of interest)?

(b) In order to obtain results with a certain a priori accuracy, how accurately should the data be known?

These questions are easy to discuss with the paradigm of the above model problem. In this case, we can easily compute the radius of the solution set,

$$r := \sup_{\alpha \in [\alpha_\circ - \delta, \alpha_\circ + \delta]} \frac{|x_{\alpha_\circ} - x_\alpha|}{|x_{\alpha_\circ}|}.$$

In the numerical tests, we select two different right–hand sides, namely

$$f^{(1)} = \tfrac{1}{\sqrt{5}}[2, -1]^T \notin \mathcal{N}(T) \quad \text{and} \quad f^{(2)} = \tfrac{1}{\sqrt{2}}[1, 1]^T \in \mathcal{N}(T).$$

TABLE 8 The radius of the solution set if $\delta = 0.01$ for different $\alpha_\circ$ and f.

	$\alpha_\circ$	δ	r
$f^{(1)} \notin \mathcal{N}(T)$	0.50	0.01	0.0204
	0.80	0.01	0.0526
	0.95	0.01	0.2500
$f^{(2)} \in \mathcal{N}(T)$	0.50	0.01	0.0040
	0.80	0.01	0.0046
	0.95	0.01	0.0049

TABLE 9 The accuracy δ required to obtain $r = 0.05$ for different $\alpha_\circ$.

	$\alpha_\circ$	δ	r
$f^{(1)} \notin \mathcal{N}(T)$	0.50	0.0235	0.05
	0.80	0.0094	0.05
	0.95	0.0024	0.05
$f^{(2)} \in \mathcal{N}(T)$	0.50	0.1189	0.05
	0.80	0.1046	0.05
	0.95	0.0499	0.025

Table 8 presents the radius for various values of $\alpha_\circ$ and given right–hand sides. From the results it is obvious that the difference between the two cases is very large. If $f = f^{(1)}$, then for $\alpha = 0.95 \pm 0.01$ we have $r = 25\%$, which means that any quantitative analysis is rather meaningless. If $f = f^{(2)}$, then $r < 0.5\%$ in all experiments.

Table 9 shows how accurately δ should be known in order to obtain the solution set with a radius less than 5%. Again, for $f = f^{(1)}$ we see that 5 % accuracy demands very accurate knowledge of α. For $f = f^{(2)}$, the situation is drastically different. In the case $f = f^{(2)}$ and $\alpha_\circ = 0.95$, the restriction $\alpha_\circ + \delta < 1$ is met before the radius grows up to 5 %.

This example underlines the importance of an a priori qualitative analysis to study the stability of the problem and, moreover, the necessity to produce quantitative information on the effects that the incomplete data have on the solutions.

4.1.3 Example 3

In this example, we study a minimization problem with constraints. Consider the problem

$$\min_{x \in K \subset \mathbb{R}^2} Q^\alpha(x),$$

where

$$Q^\alpha(x) := y \cdot x, \quad y = \frac{1}{1-\alpha} Bx + x, \quad B = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

Hence,

$$Q^{\alpha}(x) = \frac{1}{1-\alpha} Bx \cdot x + |x|^2 = \frac{1}{1-\alpha}|Ex|^2 + |x|^2$$
$$= \frac{1}{1-\alpha}(x_1 - x_2)^2 + x_1^2 + x_2^2,$$

where

$$E = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}.$$

The set of admissible solutions is

$$K := \{x \in \mathbb{R}^2 \,|\, x_2 = ax_1 + b,\ a, b \in \mathbb{R}\}.$$

The kernel of the matrix E is the line

$$\mathcal{N}(E) = \{x \in \mathbb{R}^2 \mid x_1 = x_2\}.$$

We consider the simplest case where only the parameter α is not fully determined, but the coefficients a and b are exactly known.

The minimizer of Q^{α} in K can be easily computed. Substituting $x_2 = ax_1 + b$ to Q we obtain the form

$$\begin{aligned}
Q^{\alpha}_{|x \in K}(x_1) &= \frac{1}{1-\alpha}(x_1 - ax_1 - b)^2 + x_1^2 + (ax_1 + b)^2 \\
&= \frac{1}{1-\alpha}\left((1-a)^2 x_1^2 - 2(1-a)bx_1 + b^2\right) + x_1^2 + a^2 x_1^2 + 2abx_1 + b^2 \\
&= \left(\frac{1}{1-\alpha}(1-a)^2 + 1 + a^2\right) x_1^2 + \left(\frac{1}{1-\alpha} 2b(a-1) + 2ab\right) x_1 + \\
&\quad + \frac{1}{1-\alpha} b^2 + b^2,
\end{aligned}$$

which is to be minimized with respect to x_1. It is easy to check that the minimizer must satisfy the condition

$$\left(\frac{1}{1-\alpha}(1-a)^2 + 1 + a^2\right) x_1 + \frac{1}{1-\alpha} b(a-1) + ab = 0.$$

The minimum of Q^{α} is attained at

$$x^0 = \frac{b}{(\alpha-2)\, a^2 + 2\, a + \alpha - 2} \begin{bmatrix} -1 - a\,(\alpha - 2) \\ a + \alpha - 2 \end{bmatrix}.$$

The minimum value of Q^{α} is

$$Q^{\alpha}_0 = \frac{b^2\ (\alpha - 3)}{(\alpha-2)\ a^2 + 2\, a + \alpha - 2}.$$

Also,

$$y^0 = \frac{b\ (\alpha - 3)}{(\alpha-2)\ a^2 + 2\, a + \alpha - 2} \begin{bmatrix} -a \\ 1 \end{bmatrix}.$$

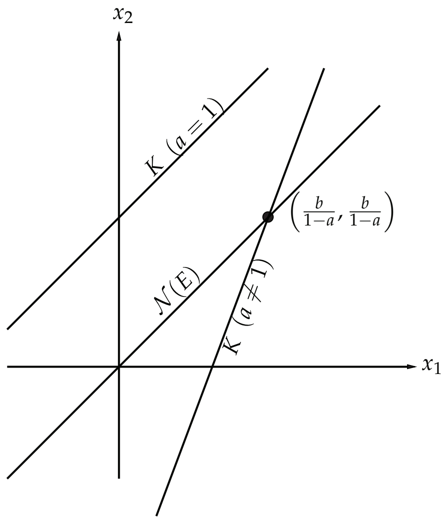

FIGURE 8 Illustration of the sets K and $\mathcal{N}(E)$.

We can compute the derivatives

$$\frac{\mathrm{d}x^0}{\mathrm{d}\alpha} = \frac{b\ (a-1)\ (a+1)}{(2\,a - 2\,a^2 + \alpha\ (a^2+1) - 2)^2} \begin{bmatrix} -1 \\ -a \end{bmatrix},$$

$$\frac{\mathrm{d}Q_0^\alpha}{\mathrm{d}\alpha} = \frac{b^2\,(a+1)^2}{(2\,a - 2\,a^2 + \alpha\ (a^2+1) - 2)^2},$$

$$\frac{\mathrm{d}y^0}{\mathrm{d}\alpha} = \frac{b\,(a+1)^2}{(2\,a - 2\,a^2 + \alpha\ (a^2+1) - 2)^2} \begin{bmatrix} -a \\ 1 \end{bmatrix}.$$

The behavior of the solution on the limit $\alpha \to 1$ depends on whether sets K and $\mathcal{N}(E)$ intersect (see Figure 8). We distinguish between two cases:

- **Case $\mathbf{a = 1}$: the sets $\mathcal{N}(E)$ and K do not have a common point.**

  Then,

  $$x^0 = \frac{b}{2} \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \quad \frac{\mathrm{d}x^0}{\mathrm{d}\alpha} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

  Note that x is unaffected by changes of α. However, for Q^α and y we have

  $$\lim_{\alpha \to 1} |Q_0^\alpha| = \infty, \quad \lim_{\alpha \to 1} |y^0| = \infty$$

  and for derivatives

  $$\lim_{\alpha \to 1} \left| \frac{\mathrm{d}Q_0^\alpha}{\mathrm{d}\alpha} \right| = \infty, \quad \lim_{\alpha \to 1} \left| \frac{\mathrm{d}y^0}{\mathrm{d}\alpha} \right| = \infty$$

  and for logarithmic derivatives

  $$\lim_{\alpha \to 1} \frac{1}{|Q_0^\alpha|} \left| \frac{\mathrm{d}Q_0^\alpha}{\mathrm{d}\alpha} \right| = \infty, \quad \lim_{\alpha \to 1} \frac{1}{|y^0|} \left| \frac{\mathrm{d}y^0}{\mathrm{d}\alpha} \right| = \infty.$$

Thus, they become very sensitive to variations of α if α is close to one.

- **Case a $\neq$ 1 : the sets $\mathcal{N}(E)$ and K have a common point.**

  Then

$$\lim_{\alpha\to 1} x^0 = \frac{b}{1-a}\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \lim_{\alpha\to 1} Q_0^\alpha = \frac{2b^2}{(1-a)^2}, \quad \lim_{\alpha\to 1} y^0 = \frac{2b}{(1-a)^2}\begin{bmatrix} -a \\ 1 \end{bmatrix},$$

  and

$$\lim_{\alpha\to 1} \frac{\mathrm{d}x^0}{\mathrm{d}\alpha} = \frac{b(a+1)^2}{(a-1)^4}\begin{bmatrix} -a \\ -1 \end{bmatrix}, \quad \lim_{\alpha\to 1} \frac{\mathrm{d}Q_0^\alpha}{\mathrm{d}\alpha} = -\frac{b^2(a+1)^2}{(1-a)^4},$$

$$\lim_{\alpha\to 1} \frac{\mathrm{d}y^0}{\mathrm{d}\alpha} = -\frac{b(a+1)^2}{(1-a)^3}\begin{bmatrix} 1 \\ a \end{bmatrix}.$$

This study indicates that for this elementary model problem different constraints have a crucial effect on the reliability of a quantitative analysis. More precisely, if α is close to one, then a quantitative analysis of the problem in the case $a = 1$ makes no sense. We emphasize this fact by considering a case where the parameter α is not fully known, but instead belongs to an interval

$$\alpha \in [\alpha_\circ - \delta, \alpha_\circ + \delta],$$

where $\alpha_\circ$ is a known mean value and δ is the magnitude of the indeterminacy. This indeterminacy of the parameter introduces a set of solutions,

$$\mathcal{S}_x := \{x \in \mathbb{R}^2 \mid x = \arg\min_{z\in K} Q^\alpha(z),\ \alpha \in [\alpha_\circ - \delta, \alpha_\circ + \delta]\}.$$

Similarly,

$$\mathcal{S}_y := \{y \in \mathbb{R}^2 \mid y = \tfrac{1}{1-\alpha} B x_\alpha + x_\alpha,\ \ x_\alpha = \arg\min_{z\in K} Q^\alpha(z),\ \alpha \in [\alpha_\circ - \delta, \alpha_\circ + \delta]\}$$

and

$$\mathcal{S}_Q := \{Q \in \mathbb{R} \mid Q = \min_{z\in K} Q^\alpha(z),\ \alpha \in [\alpha_\circ - \delta, \alpha_\circ + \delta]\}.$$

We compute the radii of the solution sets,

$$r_x := \sup_{x\in\mathcal{S}_x^K} \frac{|x_\circ - x|}{|x_\circ|},$$

$$r_y := \sup_{y\in\mathcal{S}_y^K} \frac{|y_\circ - y|}{|y_\circ|},$$

$$r_Q := \sup_{Q\in\mathcal{S}_Q^K} \frac{|Q_\circ - Q|}{|Q_\circ|},$$

where subscript $\circ$ denotes the solution corresponding to the problem where $\alpha = \alpha_\circ$.

TABLE 10 Normalized radii of the solution set for $\alpha = 0.50 \pm 0.01$.

	r_x	r_y	r_Q
$K^{(1)}$	0	0.0163	0.0163
$K^{(2)}$	0.0094	0.0104	0.0104

TABLE 11 Normalized radii of the solution set for $\alpha = 0.95 \pm 0.01$.

	r_x	r_y	r_Q
$K^{(1)}$	0	0.2439	0.2439
$K^{(2)}$	0.0385	0.0366	0.0366

In the numerical tests we consider two different constraints,

$$K^{(1)} = \{x \in \mathbb{R}^2 \mid x_2 = x_1 + 1\}$$

and

$$K^{(2)} = \{x \in \mathbb{R}^2 \mid x_2 = \tfrac{1}{2}x_1 + 1\},$$

where in particular $K^{(2)}$ has a common point with $\mathcal{N}(E)$ and $K^{(1)}$ has not. The computed radii for the two intervals are collected in Tables 10 and 11. The solution sets related to $K^{(1)}$ and $K^{(2)}$ are rather similar if $\alpha = 0.50 \pm 0.01$, but significantly different in the case $\alpha = 0.95 \pm 0.01$. In the case $K = K^{(2)}$, if α is close to one, α has to be known extremely accurately. In Table 11 the data are known at 1% accuracy, but y and Q by 24 %!

The asymptotic results derived earlier only indicate the severity of the indeterminacy. For a creditable quantitative analysis one demands quantitative information concerning the magnitude of the indeterminacy in the data and the relation between this indeterminacy and the solution set.

4.2 Estimates of worst and best case scenario errors

The elementary problems presented above give a rather obscure view of the complicated effects that may arise in problems with incompletely known data.

On the first glance, the computation of the radius of the solution set or best and worst case errors for problems generated by PDEs seems an unfeasible task, since elements of the set $\mathcal{S}(\mathcal{D})$ are generally unknown. The remedies are deviation estimates and their following properties:

(a) They are guaranteed (in the sense discussed in Theorems 2.2 and 2.3),

(b) they provide the exact error with some auxiliary functions and

(c) they depend *explicitly* on the problem data.

Due to the properties mentioned before, the corresponding majorant and minorant can be applied to compute upper and lower bounds of errors generated by the indeterminacy.

The benefits of using deviation estimates are introduced in the following Proposition, where worst and best case errors are controlled. Here, we write the majorant (minorant) as $\mathcal{M}_{\oplus}(v, y, \beta; D)$ to highlight the fact that the majorant (minorant) depends explicitly on the problem data D. We recall that the set of admissible data, i.e., data within the allowed perturbations, are denoted by $\mathcal{D}$ and the set of solutions by $\mathcal{S}(\mathcal{D})$.

Proposition 4.1. *Let $v \in V_0$ and u be the exact solution of problem* (2.35)*. Then the worst case error* (4.3) *and the best case error* (4.4) *can be controlled as follows,*

$$\underline{K} \sup_{w} \sup_{D \in \mathcal{D}} \mathcal{M}_{\ominus}(v, w; D) \leq e^2_{\text{worst}} \leq \overline{K} \inf_{y,\beta} \sup_{D \in \mathcal{D}} \mathcal{M}_{\oplus}(v, y, \beta; D)$$

and

$$\underline{K} \sup_{w} \inf_{D \in \mathcal{D}} \mathcal{M}_{\ominus}(v, w; D) \leq e^2_{\text{best}} \leq \overline{K} \inf_{y,\beta} \inf_{D \in \mathcal{D}} \mathcal{M}_{\oplus}(v, y, \beta; D),$$

where constants $\underline{K}$ and $\overline{K}$ satisfy

$$\underline{K} \, \| w \|^2_D \leq \| w \|^2_\circ \leq \overline{K} \, \| w \|^2_D, \quad \forall w \in V_0, \ D \in \mathcal{D}, \tag{4.10}$$

and where $\| \cdot \|_D$ denotes the energy norm generated by the problem data D.

Proof. We estimate the worst case error from above using the majorant (see Theorem 2.3), and inequalities (2.7) and (4.10),

$$e^2_{\text{worst}} = \sup_{u \in \mathcal{S}(\mathcal{D})} \| u - v \|^2_\circ \leq \overline{K} \sup_{D \in \mathcal{D}} \inf_{y,\beta} \mathcal{M}_{\oplus}(v, y, \beta; D) \leq \overline{K} \inf_{y,\beta} \sup_{D \in \mathcal{D}} \mathcal{M}_{\oplus}(v, y, \beta; D).$$

Alternatively, we can apply the minorant (see Theorem 2.2) and (4.10),

$$e^2_{\text{worst}} = \sup_{u \in \mathcal{S}(\mathcal{D})} \| u - v \|^2_\circ \geq \underline{K} \sup_{D \in \mathcal{D}} \sup_{w} \mathcal{M}_{\ominus}(v, w; D) = \underline{K} \sup_{w} \sup_{D \in \mathcal{D}} \mathcal{M}_{\ominus}(v, w; D).$$

The estimation of the best case error is similar,

$$e^2_{\text{best}} = \inf_{u \in \mathcal{S}(\mathcal{D})} \| u - v \|^2_\circ \leq \overline{K} \inf_{D \in \mathcal{D}} \inf_{y,\beta} \mathcal{M}_{\oplus}(v, y, \beta; D) = \overline{K} \inf_{y,\beta} \inf_{D \in \mathcal{D}} \mathcal{M}_{\oplus}(v, y, \beta; D)$$

and

$$e^2_{\text{best}} = \inf_{u \in \mathcal{S}(\mathcal{D})} \| u - v \|^2_\circ \geq \underline{K} \inf_{D \in \mathcal{D}} \sup_{w} \mathcal{M}_{\ominus}(v, w; D) \geq \underline{K} \sup_{w} \inf_{D \in \mathcal{D}} \mathcal{M}_{\ominus}(v, w; D).$$

□

The key part of this proposition is the following: Since the error functionals depend explicitly on the problem data, the extremum over the set of admissible data can in many cases be estimated or computed analytically. Thus, instead of solving the original problem numerous times, we end up with a problem of finding an extremum of a single functional. Moreover, it is not necessary to compute this extremal auxiliary function exactly, since any approximation of it will provide an error bound. Naturally, the more computational effort we invest in this task, the sharper bound we obtain.

This approach was studied first in [25]. The practical impact of Proposition 4.1 is that the ratio of e_{best} and e_{worst} can be estimated from both sides. This ratio indicates how close the approximative solution is to the solution set. If this ratio is close to one, then the magnitude of the solution set is small compared to the approximation error and it makes sense to invest more in the computation. However, if the ratio is very small, then the approximate solution is very close to (or inside of) the solution set and further adaptation of meshes or other means to improve the approximation makes no sense.

4.3 Estimates of the radius of the solution set

A more fundamental quantity to study is the radius of the solution set, which has been studied for diffusion type problems in [26, 28, 29]. The radius is not related to any particular approximation, but contains information about the effect of incomplete knowledge on the problem itself. In this Section, we derive and discuss various estimates for the radius for the generalized model (2.34) and the reaction diffusion problem (2.23–2.26).

We recall the problem (2.35): find $u \in V_0$ such that

$$(\mathcal{A}\Lambda u, \Lambda w)_U = (f, w)_{\mathcal{V}}, \quad \forall w \in V_0.$$

We assume that the operator $\mathcal{A}$ is not completely known, but it belongs to a set of admissible operators $\mathcal{D}_{\mathcal{A}}$.

$$\mathcal{D}_{\mathcal{A}} := \{\mathcal{A} \in \mathcal{L}(U, U) | \mathcal{A} = \mathcal{A}_\circ + \delta\Psi,\ \|\Psi\|_{\mathcal{L}} \leq 1,\ \delta \geq 0\}, \tag{4.11}$$

where $A_\circ$ is known a "mean" operator and $\delta \geq 0$ is the indeterminacy magnitude. We denote by $\|\Psi\|_{\mathcal{L}}$ the operator norm of Ψ, i.e.,

$$\|\Psi\|_{\mathcal{L}} = \sup_{\substack{y \in U \\ y \neq 0}} \frac{\|\Psi y\|_U}{\|y\|_U}.$$

We note that all members of the set $\mathcal{D}_{\mathcal{A}}$ generate an energy norm of their own

$$|\!|\!| w |\!|\!|_{\mathcal{A}}^2 := (\mathcal{A}\Lambda w, \Lambda w)_U. \tag{4.12}$$

In this Section, an energy norm without any subscript is defined as follows:

$$|\!|\!| w |\!|\!|^2 := (\Lambda w, \Lambda w)_U.$$

We assume that the non-perturbed operator is elliptic and bounded,

$$\underline{c}\|y\|_U^2 \le (\mathcal{A}_\circ y, y)_U \le \overline{c}\|y\|_U^2, \quad \forall\, y \in U. \tag{4.13}$$

The perturbations are restricted by

$$\theta = \frac{\delta}{\underline{c}} < 1$$

to guarantee the ellipticity of perturbed problems. This normalization was motivated by the Example 1 on page 61. We recall also the condition number of $\mathcal{A}_\circ$,

$$\mathrm{cond}(\mathcal{A}_\circ) = \frac{\overline{c}}{\underline{c}}.$$

Of special interest is the energy norm generated by the known "mean" operator $|||\cdot|||_{\mathcal{A}_\circ}$. The constants of the norm equality between $|||\, w\, |||_{\mathcal{A}_\circ}$ and arbitrary $|||\, w\, |||_{\mathcal{A}}$, where $\mathcal{A} \in \mathcal{D}_{\mathcal{A}}$, play an important role in the subsequent analysis.

Proposition 4.2. *Let the set D_A be defined as in* (4.11)*, where $A_\circ$ satisfies* (4.13)*. Then the energy norms related to $\mathcal{A}_\circ$ and any $\mathcal{A} \in D_A$ are equal,*

$$\underline{K}\, |||\, w\, |||_{\mathcal{A}}^2 \le |||\, w\, |||_{\mathcal{A}_\circ}^2 \le \overline{K}\, |||\, w\, |||_{\mathcal{A}}^2, \quad \forall\, w \in V,\ \mathcal{A} \in D_A,$$

where

$$\underline{K} := \max\left\{\frac{1}{\mathrm{cond}(\mathcal{A}_\circ)+\theta}, \frac{1-2\theta}{1-\theta}\right\} \quad \textit{and} \quad \overline{K} := \frac{1}{1-\theta}.$$

Proof. We note that

$$(\underline{c}-\delta)\|y\|_U^2 \le (\mathcal{A}y, y)_U \le (\overline{c}+\delta)\|y\|_U^2, \quad \forall\, y \in U,\ \mathcal{A} \in D_{\mathcal{A}}.$$

By the definition (4.11),

$$|||\, w\, |||_{\mathcal{A}_\circ}^2 = |||\, w\, |||_{\mathcal{A}}^2 - \delta(\Psi\Lambda w, \Lambda w)_U.$$

From this follows by standard inequalities and assumption over Ψ,

$$|||\, w\, |||_{\mathcal{A}_\circ}^2 \le |||\, w\, |||_{\mathcal{A}}^2 + \delta\|\Psi\Lambda w\|_U\, \|\Lambda w\|_U \le |||\, w\, |||_{\mathcal{A}}^2 + \delta\, |||\, w\, |||^2 \le \left(1 + \frac{\delta}{\underline{c}-\delta}\right) |||\, w\, |||_{\mathcal{A}}^2\,.$$

Similarly for lower bound,

$$|||\, w\, |||_{\mathcal{A}_\circ}^2 \ge |||\, w\, |||_{\mathcal{A}}^2 - \delta\|\Psi\Lambda w\|_U\, \|\Lambda w\|_U \ge |||\, w\, |||_{\mathcal{A}}^2 - \delta\, |||\, w\, |||^2 \ge \left(1 - \frac{\delta}{\underline{c}-\delta}\right) |||\, w\, |||_{\mathcal{A}}^2\,.$$

An alternative lower bound is

$$|||\, w\, |||_{\mathcal{A}_\circ}^2 \ge \underline{c}\, |||\, w\, |||^2 \ge \frac{\underline{c}}{\overline{c}+\delta}\, |||\, w\, |||_{\mathcal{A}}^2\,.$$

Clearly, the maximum of lower bounds is also a lower bound. The definition of θ leads to the statement. $\square$

The normalized radius can be estimated from both sides using only the data of the mean operator and the perturbation magnitude.

Theorem 4.1. *The radius of the solution set for the problem* (2.35)*, where* $\mathcal{A} \in \mathcal{D}_{\mathcal{A}}$*, can be estimated as follows:*

$$\sqrt{\underline{K}}\,\frac{\operatorname{cond}^{-1}(\mathcal{A}_\circ)\,\theta}{\sqrt{1-\operatorname{cond}^{-1}(\mathcal{A}_\circ)\,\theta}} \le \bar{r} \le \sqrt{\overline{K}}\,\frac{\theta}{\sqrt{1-\theta}},$$

where $\underline{K}$ *and* $\overline{K}$ *are constants from Proposition 4.2.*

Proof. First, we construct the lower bound. We can apply Proposition 4.2 and Theorem 2.2 to estimate from below

$$r^2 \ge \underline{K} \sup_{u\in\mathcal{S}(\mathcal{D})} \interleave u-u_\circ \interleave^2_{\mathcal{A}} = \underline{K} \sup_{\mathcal{A}\in\mathcal{D}_{\mathcal{A}}} \sup_{w\in V_0} \mathcal{M}^{\mathcal{A}}_{\ominus}(u_\circ,w) = \underline{K} \sup_{w\in V_0} \sup_{\mathcal{A}\in\mathcal{D}_{\mathcal{A}}} \mathcal{M}^{\mathcal{A}}_{\ominus}(u_\circ,w). \tag{4.14}$$

The minorant can be written as

$$\begin{aligned} M_\ominus(u_\circ,w) &= -((A_\circ+\delta\Psi)\Lambda w,\Lambda w)_U - 2((A_\circ+\delta\Psi)\Lambda u_\circ,\Lambda w)_U + 2(f,w)_{\mathcal{V}} \\ &= -\interleave w \interleave^2_{A_\circ} - \delta(\Psi\Lambda(w+2u_\circ),\Lambda w)_U - 2((A_\circ\Lambda u_\circ,\Lambda w)_U - (f,w)_{\mathcal{V}}), \end{aligned} \tag{4.15}$$

where the last term vanishes due to the Galerkin orthogonality. We estimate $\sup_{w\in V_0} \sup_{\mathcal{A}\in\mathcal{D}_{\mathcal{A}}} M^{\mathcal{A}}_{\ominus}(u_\circ,w)$ from below and set $w := \alpha u_\circ$,

$$M_\ominus(u_\circ,\alpha u_\circ) = -\alpha^2 \interleave u_\circ \interleave^2_{A_\circ} - \delta(\alpha+2)\alpha(\Psi\Lambda u_\circ,\Lambda u_\circ)_U.$$

Taking supremum ($\alpha > 0$) leads to

$$\begin{aligned} \sup_{\|\Psi\|_{\mathcal{L}}\le 1} M_\ominus(u_\circ,\alpha u_\circ) &= -\alpha^2 \interleave u_\circ \interleave^2_{A_\circ} + \delta(\alpha+2)\alpha \interleave u_\circ \interleave^2 \\ &= \alpha(2\delta \interleave u_\circ \interleave^2 + \alpha(\delta \interleave u_\circ \interleave^2 - \interleave u_\circ \interleave^2_{A_\circ})). \end{aligned} \tag{4.16}$$

The expression attains a maximum value with

$$\tilde{\alpha} := \frac{\delta \interleave u_\circ \interleave^2}{\interleave u_\circ \interleave^2_{A_\circ} - \delta \interleave u_\circ \interleave^2}.$$

Substituting this yields

$$\sup_{\|\Psi\|_{\mathcal{L}}\le 1} M_\ominus(u_\circ,\tilde{\alpha} u_\circ) = \frac{\delta^2 \interleave u_\circ \interleave^4}{\interleave u_\circ \interleave^2_{\mathcal{A}_\circ} - \delta \interleave u_\circ \interleave^2} \ge \frac{\left(\frac{\delta}{\underline{\underline{c}}}\right)^2}{1-\left(\frac{\delta}{\underline{\underline{c}}}\right)} \interleave u_\circ \interleave^2_{\mathcal{A}_\circ}.$$

Substituting this expression to (4.14), dividing by $\interleave u_\circ \interleave^2_{A_\circ}$, and taking a square root leads to the lower bound.

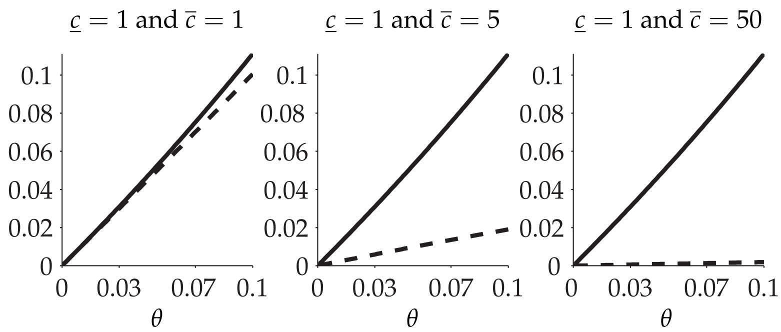

FIGURE 9 Lower (dashed line) and upper (solid line) bounds from Theorem 4.1 with different spectral ranges for the non–perturbed "mean" matrix.

Similarly, we estimate from above and apply (2.7),

$$r^2 \leq \overline{K} \sup_{u \in \mathcal{S}(\mathcal{D})} \||\, u - u_\circ \,\||^2_{\mathcal{A}} = \overline{K} \sup_{\mathcal{A} \in \mathcal{D}_\mathcal{A}} \inf_{\substack{y \in Q \\ \beta > 0}} \mathcal{M}^{\mathcal{A}}_{\oplus}(u_\circ, y, \beta) \leq \overline{K} \inf_{\substack{y \in Q \\ \beta > 0}} \sup_{\mathcal{A} \in \mathcal{D}_\mathcal{A}} \mathcal{M}^{\mathcal{A}}_{\oplus}(u_\circ, y, \beta). \tag{4.17}$$

We estimate infimum from above and set $y := \mathcal{A}_\circ \Lambda u_\circ$. Then the majorant can be written as

$$M^{\mathcal{A}}_{\oplus}(u_\circ, \mathcal{A}_\circ \Lambda u_\circ, \beta) = (1+\beta)(\Lambda u_\circ - \mathcal{A}^{-1}\mathcal{A}_\circ \Lambda u_\circ, \mathcal{A}\Lambda u_\circ - \mathcal{A}_\circ \Lambda u_\circ)_U + \\ + \left(1 + \frac{1}{\beta}\right) \frac{C_F^2}{\underline{c} - \delta} \| f - \Lambda^* \mathcal{A}_\circ \Lambda u_\circ \|^2_{\mathcal{V}}, \tag{4.18}$$

where the last term vanishes, since the exact flux satisfies the equilibrium condition (2.41) and we can take β arbitrarily close to zero. Thus

$$M^{\mathcal{A}}_{\oplus}(u_\circ, \mathcal{A}_\circ \Lambda u_\circ, 0) = (\mathcal{A}^{-1}((\mathcal{A}_\circ + \delta\Psi)\Lambda u_\circ - \mathcal{A}_\circ \Lambda u_\circ), (\mathcal{A}_\circ + \delta\Psi)\Lambda u_\circ - \mathcal{A}_\circ \Lambda u_\circ)_U \\ \leq \frac{\delta^2}{\underline{c} - \delta}(\Psi \Lambda u_\circ, \Psi \Lambda u_\circ)_U \leq \frac{\delta^2}{\underline{c} - \delta} \||\, u_\circ \,\||^2 \leq \frac{\delta^2}{\underline{c}(\underline{c} - \delta)} \||\, u_\circ \,\||^2_{\mathcal{A}_\circ} = \frac{\theta^2}{1 - \theta} \||\, u_\circ \,\||^2_{\mathcal{A}_\circ} . \tag{4.19}$$

Substituting this estimate to (4.17), dividing by $\||\, u_\circ \,\||^2_{\mathcal{A}_\circ}$ and taking a square root leads to the upper bound. □

The accuracy of the estimates in Theorem 4.1 depends on estimate (4.13). In Figure 9, we have plotted upper and lower bounds for the radius from Theorem 4.1. Obviously, the accuracy of the lower bound deteriorates quickly as the spectral range of the operator $\mathcal{A}_\circ$ increases. Of course this does not mean that problems where the operator has large spectral range are less sensitive to the indeterminacy. The applied technique simply does not provide an efficient lower bound for the radius. Next, we discuss the reason for this.

Let $\tilde{u}$ be the most distant member of u in the solution set, i.e.,

$$\||\, u_\circ - \tilde{u} \,\||_{\mathcal{A}_\circ} = \sup_{u \in \mathcal{S}(\mathcal{D})} \||\, u_\circ - u \,\||_{\mathcal{A}_\circ} .$$

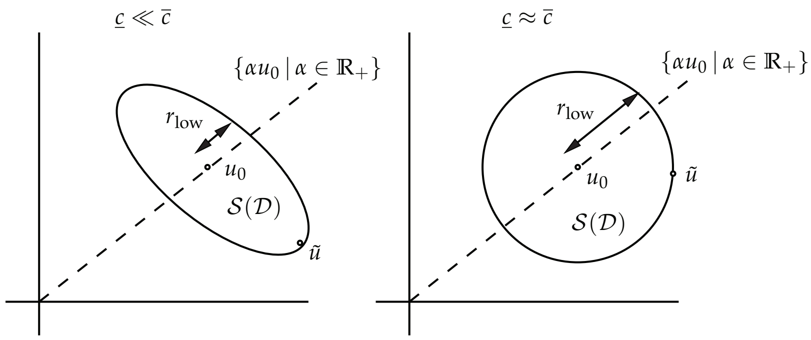

FIGURE 10 Illustration of the efficiency of selecting an auxiliary function w to construct the lower bound for the radius. r_{low} refers to the lower estimate of Theorem 4.1.

Then

$$\hat{r} = \frac{|||\ u_\circ - \tilde{u}\ |||_{\mathcal{A}_\circ}}{|||\ u_\circ\ |||_{\mathcal{A}_\circ}}.$$

When we derived the lower bound in the proof of Theorem 4.1, we estimated in (4.14) as follows:

$$r^2 = |||\ u_\circ - \tilde{u}\ |||^2_{\mathcal{A}_\circ} \geq \underline{K}\ |||\ u_\circ - \tilde{u}\ |||^2_{\tilde{\mathcal{A}}} = \sup_{w \in V_0} \mathcal{M}^{\tilde{\mathcal{A}}}_{\ominus}(u_\circ, w).$$

Compared to (4.14), due to the fact that we know $\tilde{u}$ and the respective operator $\tilde{\mathcal{A}}$, we were not required to take supremum over the solution set. Recall from Theorem 2.2 that the supremum is attained if $w := u_\circ - \tilde{u}$. In the derivation of the lower bound, we estimated the supremum from below and selected $w := \alpha u_\circ$. This choice (and lower estimate) is efficient if $\alpha u_\circ \approx u_\circ - \tilde{u}$. Whether this happens or not depends on the "shape" of the solution set. We illustrate this phenomenon in Figure 10.

However, the upper bound is sharp. In the following section, we construct an example that admits an analytical solution and compute the radius of the solution set for a certain type of perturbation. Then we compare this true radius and the upper bound from Theorem 4.1.

Example: The upper bound in Theorem 4.1 is attainable

Bounds in Theorem 4.1 are derived for the generalized problem (2.35). These bounds are valid for any right–hand side f and differential operator Λ with operator $\mathcal{A}_\circ$ satisfying solvability conditions, and they can be computed a priori. In this section, we present an elementary diffusion problem for which we can compute the radius of the solution set. Then we show that with some problem parameters the radius attains the upper bound derived in Theorem 4.1, i.e., there is no gap between the true radius and the estimate.

Consider the stationary heat equation problem in the unit square,

$$\begin{aligned} -\operatorname{div}(A\nabla u) &= f, \quad \text{if } (x,y) \in \Omega := [0,1] \times [0,1], \\ u &= 0, \quad \text{if } (x,y) \in \Gamma, \end{aligned}$$

where f coincides with the eigenfunction, i.e.,

$$f(x,y) := \sin(k_1 \pi x) \sin(k_2 \pi y), \quad \text{where } k_1, k_2 \in \mathbb{N} \tag{4.20}$$

Let

$$A_\circ := \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix},$$

where coefficients a_1 and a_2 are positive constants. In this case

$$u_\circ(x,y) = \frac{\sin(k_1 \pi x) \sin(k_2 \pi y)}{\pi^2 (a_1 k_1^2 + a_2 k_2^2)} = \frac{f(x,y)}{\pi^2 \bar{a} \cdot \bar{k}},$$

where for the sake of convenience we use the notation

$$\bar{a} := \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \quad \text{and} \quad \bar{k} := \begin{bmatrix} k_1^2 \\ k_2^2 \end{bmatrix}.$$

Consider the same problem with the perturbed matrix,

$$\tilde{A} = A_\circ + \delta \Psi = \begin{bmatrix} a_1 + \delta \epsilon_1 & 0 \\ 0 & a_2 + \delta \epsilon_2 \end{bmatrix},$$

i.e., perturbations are generated by a diagonal constant matrix. Then the restriction $|\Psi| \leq 1$ leads to the condition

$$\bar{\epsilon} \in \mathcal{E} := \{\epsilon_1^2 + \epsilon_2^2 \leq 1\}, \tag{4.21}$$

where

$$\bar{\epsilon} := \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}.$$

The solution of our perturbed problem is given by the relation

$$\tilde{u}(x,y) = \frac{f(x,y)}{\pi^2 (\bar{a} + \delta \bar{\epsilon}) \cdot \bar{k}}.$$

Now, we can observe how the perturbation of the matrix affects the solution:

$$\frac{\|\| u_\circ - \tilde{u} \|\|}{\|\| u_\circ \|\|} = \frac{|\delta \bar{\epsilon} \cdot \bar{k}|}{|\bar{a} \cdot \bar{k} + \delta \bar{\epsilon} \cdot \bar{k}|}. \tag{4.22}$$

It is easy to see that

$$\max_{\bar{\epsilon} \in \mathcal{E}} \frac{|\delta \bar{\epsilon} \cdot \bar{k}|}{|\bar{a} \cdot \bar{k} + \delta \bar{\epsilon} \cdot \bar{k}|} = \frac{\delta}{\frac{\bar{a} \cdot \bar{k}}{|\bar{k}|} - \delta}. \tag{4.23}$$

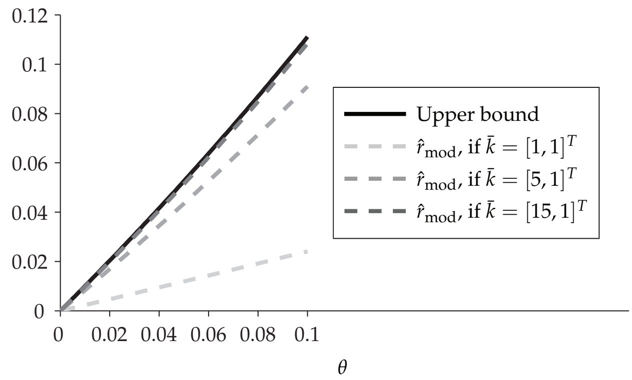

FIGURE 11 Comparison of the true radius r_{mod} with different values k_1 and k_2, and the upper bound for the radius in Theorem 4.1.

The maximal value in (4.23) is attained if

$$\bar{\epsilon} = -\frac{\bar{k}}{|\bar{k}|}.$$

Without a loss of generality we assume $a_1 \leq a_2$. In this case the normalized perturbation is

$$\theta = \frac{\delta}{a_1}.$$

Thus, for this model problem, the radius of the solution set is given by the relation

$$\hat{r}_{\text{mod}} := \max_{\bar{\epsilon} \in \mathcal{E}} \frac{\| u_\circ - \tilde{u} \|}{\| u_\circ \|} = \frac{\theta}{\frac{\bar{a} \cdot \bar{k}}{a_1 |\bar{k}|} - \theta}. \tag{4.24}$$

Remark 4.1. *If in* (4.24) *we let k_1 tend to infinity, then the ratio*

$$\frac{\frac{1}{a_1} \bar{a} \cdot \bar{k}}{|\bar{k}|} = \frac{k_1^2 + \text{cond}(A_\circ) k_2^2}{\sqrt{k_1^4 + k_2^4}}$$

tends to one and the right–hand side of (4.24) *tends to the upper bound established in Theorem 4.1. This means that the upper bound in Theorem 4.1 indeed provides a sharp estimate of the radius and cannot be improved.*

In Figure 11 we plot r_{mod} with different values k_1 and k_2 in (4.20) and compare them to the upper bound established in Theorem 4.1. Clearly, as the value of k_1 increases, the radius approaches the upper bound derived in Theorem 4.1.

Incompletely known right–hand side

The estimates of Theorem 4.1 can be extended if we additionally assume that the right–hand side of the equation is not fully determined either, but that f belongs to the set

$$\mathcal{D}_f := \{f \in \mathcal{V} |\ f = f_\circ + \xi\},$$

where

$$\|\xi\|_{\mathcal{V}} \leq \epsilon.$$

Unfortunately these bounds are not explicit, but depend on the "mean" solution $u_\circ$.

Theorem 4.2. *The radius of the solution set for the problem* (2.35), *where $\mathcal{A}$ is defined as in* (4.11) *and $f \in \mathcal{D}_f$ can be estimated as follows:*

$$\sqrt{\underline{K}}\ \frac{\sqrt{\mathrm{cond}^{-1}(\mathcal{A}_\circ)\theta \ |\!|\!|\ u_\circ \ |\!|\!|^2_{\mathcal{A}_\circ} + \epsilon \|u_\circ\|_{\mathcal{V}}}}{\sqrt{1 - \mathrm{cond}^{-1}(\mathcal{A}_\circ)\theta}} \leq r \leq \sqrt{\overline{K}}\ \frac{\theta \ |\!|\!|\ u_\circ \ |\!|\!|_{\mathcal{A}_\circ} + \frac{C_F}{\sqrt{\underline{c}}}\epsilon}{\sqrt{1-\theta}},$$

where constants $\underline{K}$ and $\overline{K}$ are defined in Proposition 4.2, $\underline{c}$ and $\overline{c}$ are from the inequality (4.13) *and C_F and c_1 are from* (2.36) *and* (2.38).

Proof. The proof is very similar to that of Theorem 4.1. First, we construct the lower bound. The key step is again to apply Proposition 4.2 and Theorem 2.2 to estimate from below,

$$r^2 \geq \underline{K} \sup_{u \in \mathcal{S}(\mathcal{D})} |\!|\!|\ u - u_\circ \ |\!|\!|^2_{\mathcal{A}} = \underline{K} \sup_{\substack{\mathcal{A} \in \mathcal{D}_{\mathcal{A}} \\ f \in \mathcal{D}_f}} \sup_{w \in V_0} \mathcal{M}^{\mathcal{A}}_{\ominus}(u_\circ, w) = \underline{K} \sup_{w \in V_0} \sup_{\substack{\mathcal{A} \in \mathcal{D}_{\mathcal{A}} \\ f \in \mathcal{D}_f}} \mathcal{M}^{\mathcal{A}}_{\ominus}(u_\circ, w) \tag{4.25}$$

$$\begin{aligned} M^{\mathcal{A}}_{\ominus}(u_\circ, w) &= -((A_\circ + \delta\Psi)\Lambda w, \Lambda w)_U - 2((A_\circ + \delta\Psi)\Lambda u_\circ, \Lambda w)_U + 2(f_\circ + \xi, w)_{\mathcal{V}} \\ &= - |\!|\!|\ w \ |\!|\!|^2_{A_\circ} - \delta(\Psi\Lambda(w + 2u_\circ), \Lambda w)_U + (\xi, w)_{\mathcal{V}} - 2((A_\circ \Lambda u_\circ, \Lambda w)_U - (f_\circ, w)_{\mathcal{V}}), \end{aligned}$$

where the last term vanishes due to the Galerkin orthogonality. We estimate $\sup_{w \in V_0} \sup_{(A,l) \in D} M_{\ominus}(u_\circ, w)$ from below and set $w := \alpha u_\circ$,

$$M_{\ominus}(u_\circ, \alpha u_\circ) = -\alpha^2 \ |\!|\!|\ u_\circ \ |\!|\!|^2_{A_\circ} - \delta(\alpha + 2)\alpha(\Psi\Lambda u_\circ . \Lambda u_\circ)_U + 2\alpha(\xi, u_\circ)_{\mathcal{V}}.$$

Taking supremum over the uncertain perturbations ($\alpha > 0$) leads to

$$\begin{aligned} \sup_{\|\Psi\|_{\mathcal{L}} \leq 1,\ \|\xi\|_{\mathcal{V}} \leq \epsilon} M_{\ominus}(u_\circ, \alpha u_\circ) &= -\alpha^2 \ |\!|\!|\ u_\circ \ |\!|\!|^2_{A_\circ} + \delta(\alpha+2)\alpha \ |\!|\!|\ u_\circ \ |\!|\!|^2 + 2\alpha\epsilon \|u_\circ\|_{\mathcal{V}} \\ &= \alpha(2(\delta \ |\!|\!|\ u_\circ \ |\!|\!|^2 + \epsilon\|u_\circ\|_{\mathcal{V}}) + \alpha(\delta \ |\!|\!|\ u_\circ \ |\!|\!|^2 - |\!|\!|\ u_\circ \ |\!|\!|^2_{\mathcal{A}_\circ})). \end{aligned} \tag{4.26}$$

The expression attains a maximum value with

$$\tilde{\alpha} := \frac{\delta \ |\!|\!|\ u_\circ \ |\!|\!|^2 + \epsilon \|u_\circ\|_{\mathcal{V}}}{|\!|\!|\ u_\circ \ |\!|\!|^2_{A_\circ} - \delta \ |\!|\!|\ u_\circ \ |\!|\!|^2}.$$

Substituting this yields

$$\sup_{\|\Psi\|_{\mathcal{L}} \leq 1,\ \|\xi\|_{\mathcal{V}} \leq \epsilon} M_{\ominus}(u_\circ, \tilde{\alpha} u_\circ) = \frac{(\delta \ |\!|\!|\ u_\circ \ |\!|\!|^2 + \epsilon\|u_\circ\|_{\mathcal{V}})^2}{|\!|\!|\ u_\circ \ |\!|\!|^2_{A_\circ} - \delta \ |\!|\!|\ u_\circ \ |\!|\!|^2} \geq \frac{(\frac{\delta}{\overline{c}} \ |\!|\!|\ u_\circ \ |\!|\!|^2_{\mathcal{A}_\circ} + \epsilon\|u_\circ\|_{\mathcal{V}})^2}{1 - \frac{\delta}{\overline{c}}},$$

where we can write $\frac{\delta}{\underline{c}} = \mathrm{cond}^{-1}(\mathcal{A}_\circ)\theta$. Substituting this to estimate (4.25) yields the lower bound.

Next, we construct the upper bound. By Proposition 4.2 and Theorem 2.3,

$$r^2 \leq \overline{K} \sup_{u\in\mathcal{S}(\mathcal{D})} \| u - u_\circ \|_{\mathcal{A}}^2 = \overline{K} \sup_{\substack{\mathcal{A}\in\mathcal{D}_\mathcal{A}\\ f\in\mathcal{D}_f}} \inf_{\substack{y\in Q\\ \beta>0}} \mathcal{M}_\oplus^\mathcal{A}(u_\circ, y, \beta) = \overline{K} \inf_{\substack{y\in Q\\ \beta>0}} \sup_{\substack{\mathcal{A}\in\mathcal{D}_\mathcal{A}\\ f\in\mathcal{D}_f}} \mathcal{M}_\oplus^\mathcal{A}(u_\circ, y, \beta). \tag{4.27}$$

The majorant is

$$\begin{aligned}\mathcal{M}_\oplus(u_\circ, y, \beta) &:= (1+\beta)(\mathcal{A}\Lambda u_\circ - y, \Lambda u_\circ - \mathcal{A}^{-1}y)_U + \left(1+\frac{1}{\beta}\right)\frac{C_F^2}{\underline{c}-\delta}\|f - \Lambda^* y\|_\mathcal{V}^2 \\ &= (1+\beta)(\mathcal{A}^{-1}(\mathcal{A}\Lambda u_\circ - y), \mathcal{A}\Lambda u_\circ - y)_U + \left(1+\frac{1}{\beta}\right)\frac{C_F^2}{\underline{c}-\delta}\|f - \Lambda^* y\|_\mathcal{V}^2. \end{aligned} \tag{4.28}$$

We substitute $\mathcal{A} = \mathcal{A}_\circ + \delta\Psi$, $f = f_\circ + \xi$ and $y = \mathcal{A}_\circ\Lambda u_\circ$ and apply standard inequalities,

$$\begin{aligned}\mathcal{M}_\oplus(u_\circ, \mathcal{A}_\circ\Lambda u, \beta) &= (1+\beta)(\mathcal{A}^{-1}(\mathcal{A}\Lambda u_\circ - \mathcal{A}_\circ\Lambda u_\circ), \mathcal{A}\Lambda u_\circ - \mathcal{A}_\circ\Lambda u_\circ)_U + \\ &\quad + \left(1+\frac{1}{\beta}\right)\frac{C_F^2}{\underline{c}-\delta}\|f_\circ - \Lambda^*\mathcal{A}_\circ\Lambda u_\circ + \xi\|_\mathcal{V}^2 \\ &\leq \frac{\delta^2}{\underline{c}-\delta}\|(\mathcal{A}_\circ + \delta\Psi)\Lambda u_\circ - \mathcal{A}_\circ\Lambda u_\circ\|_U^2 + \left(1+\frac{1}{\beta}\right)\frac{C_F^2}{\underline{c}-\delta}\|\xi\|_\mathcal{V}^2 \\ &= \frac{1}{\underline{c}-\delta}\left((1+\beta)\delta^2\|\Psi\Lambda u_\circ\|_U^2 + \left(1+\frac{1}{\beta}\right)C_F^2\|\xi\|_\mathcal{V}^2\right).\end{aligned}$$

We take supremum and estimate from above,

$$\begin{aligned}\sup_{\|\Psi\|_\mathcal{L}\leq 1,\ \|\xi\|_\mathcal{V}\leq\epsilon} \mathcal{M}_\oplus(u_\circ, \mathcal{A}_\circ\Lambda u, \beta) &\leq \frac{1}{\underline{c}-\delta}\left((1+\beta)\delta^2\|\Lambda u_\circ\|_U^2 + \left(1+\frac{1}{\beta}\right)C_F^2\epsilon^2\right) \\ &\leq \frac{1}{1-\theta}\left((1+\beta)\theta^2 \| u_\circ \|_{\mathcal{A}_\circ}^2 + \left(1+\frac{1}{\beta}\right)\frac{C_F^2}{\underline{c}}\epsilon^2\right).\end{aligned} \tag{4.29}$$

We can substitute

$$\beta := \frac{\frac{C_F}{\sqrt{\underline{c}}}\epsilon}{\theta \| u_\circ \|_{\mathcal{A}_\circ}}$$

and arrive at

$$\sup_{\|\Psi\|_\mathcal{L}\leq 1,\ \|\xi\|_\mathcal{V}\leq\epsilon} \mathcal{M}_\oplus(u_\circ, \mathcal{A}_\circ\Lambda u, \beta) \leq \frac{1}{1-\theta}\left(\theta \| u_\circ \|_{\mathcal{A}_\circ} + \frac{C_F}{\sqrt{\underline{c}}}\epsilon\right)^2.$$

We use this to estimate the right–hand side of (4.27) from above and arrive at the statement. □

4.3.1 The reaction diffusion problem

The reaction diffusion problem (2.23–2.26) has several functions depending on the material properties, namely A, ρ, and α. The case where all of them are incompletely known is discussed here. These results were presented in [28].

We assume

$$\mathcal{D}_A := \{A \in L_\infty(\Omega, \mathbb{M}^{d\times d}) \mid A = A_\circ + \delta_1\Psi,\ ||\Psi||_{L_\infty(\Omega,\mathbb{M}^{d\times d})} \leq 1\} \tag{4.30}$$

$$\mathcal{D}_\rho := \{\rho \in L_\infty(\Omega, \mathbb{R}) \mid \rho = \rho_\circ + \delta_2\psi_\rho,\ ||\psi_\rho||_{L_\infty(\Omega)} \leq 1\} \tag{4.31}$$

$$\mathcal{D}_\alpha := \{\alpha \in L_\infty(\Gamma_3, \mathbb{R}) \mid \alpha = \alpha_\circ + \delta_3\psi_\alpha,\ ||\psi_\alpha||_{L_\infty(\Gamma_3)} \leq 1\}. \tag{4.32}$$

In other words, we assume that the sets of admissible data are formed by some (limited) variations of some known "mean" data (which are denoted by subindex 0). The parameters δ_i, $i = 1,2,3$ represent the magnitude of these variations. Thus, in the case considered, the set of indeterminate data is

$$\mathcal{D} := \mathcal{D}_A \times \mathcal{D}_\rho \times \mathcal{D}_\alpha.$$

We assume that

$$\begin{array}{rcll} \underline{c}_1|\xi|^2 \leq & A_\circ\xi\cdot\xi & \leq \overline{c}_1|\xi|^2 \quad \forall \xi \in \mathbb{R}^d & \text{on } \Omega, \\ \underline{c}_2 \leq & \rho_\circ & \leq \overline{c}_2 & \text{on } \Omega, \\ \underline{c}_3 \leq & \alpha_\circ & \leq \overline{c}_3 & \text{on } \Gamma_R, \end{array}$$

where $\underline{c}_i > 0$. In view of the above-stated conditions, the "mean" problem is evidently elliptic and has a unique solution $u_\circ$. The condition

$$0 \leq \delta_i < \underline{c}_i, \qquad i = 1,2,3$$

guarantees that the perturbed problem remains elliptic and possesses a unique solution u. We define normalized perturbations and "condition numbers",

$$\theta_i := \frac{\delta_i}{\underline{c}_i} \quad \text{and} \quad \text{cond}_i := \frac{\overline{c}_i}{\underline{c}_i}, \qquad i = 1,2,3.$$

We recall the energy norm of the problem,

$$|||\, w \,|||^2_{(A,\rho,\alpha)} := \int\limits_\Omega \left(A\nabla w \cdot \nabla w + \rho w^2 \right) \mathrm{dx} + \int\limits_{\Gamma_R} \alpha w^2 \,\mathrm{ds}.$$

Occasionally, we denote $|||\, w \,|||_\circ := |||\, w \,|||_{(A_\circ,\rho_\circ,\alpha_\circ)}$. We establish the equivalence between the norms generated by the perturbed and original data.

Proposition 4.3. *Let A, ρ, and α be defined by* (4.30), (4.31) *and* (4.32) *respectively. Then*

$$\underline{C}\, |||\, w \,|||^2_{(A,\rho,\alpha)} \leq |||\, w \,|||^2_{(A_\circ,\rho_\circ,\alpha_\circ)} \leq \overline{C}\, |||\, w \,|||^2_{(A,\rho,\alpha)}, \quad \forall w \in V_0,$$

where

$$\underline{C} := \max\left\{ \min_{i\in\{1,2,3\}} \frac{1}{\text{cond}_i + \theta_i}, \min_{i\in\{1,2,3\}} \frac{1-2\theta_i}{1-\theta_i} \right\}$$

and

$$\overline{C} := \max_{i\in\{1,2,3\}} \frac{1}{1-\theta_i}.$$

Proof. We note that

$$\begin{aligned}
(\underline{c}_1-\delta_1)\ \|\!| w \|\!|^2 &\leq \|\!| w \|\!|_A^2 \leq (\overline{c}_1+\delta_1)\ \|\!| w \|\!|^2, \quad \forall\, w\in V_0,\ A\in\mathcal{D}_A.\\
(\underline{c}_2-\delta_2)\ \|\!| w \|\!|^2 &\leq \|\!| w \|\!|_\rho^2 \leq (\overline{c}_2+\delta_2)\ \|\!| w \|\!|^2, \quad \forall\, w\in V_0,\ \rho\in\mathcal{D}_\rho.\\
(\underline{c}_3-\delta_3)\ \|\!| y \|\!|^2 &\leq \|\!| w \|\!|_\alpha^2 \leq (\overline{c}_3+\delta_3)\ \|\!| y \|\!|^2, \quad \forall\, w\in V_0,\ \alpha\in\mathcal{D}_\alpha.
\end{aligned}$$

By the definitions (4.30), (4.31), and (4.32), we have

$$\|\!| w \|\!|^2_{(A_\circ,\rho_\circ,\alpha_\circ)} = \|\!| w \|\!|^2_{(A,\rho,\alpha)} - \int\limits_\Omega \left(\delta_1\Psi\nabla w\cdot\nabla w+\delta_2\psi_\rho w^2\right)\,\mathrm{dx} - \delta_3\int\limits_{\Gamma_R}\psi_\alpha w^2\,\mathrm{ds}.$$

From it follows

$$\begin{aligned}
\|\!| w \|\!|^2_{(A_\circ,\rho_\circ,\alpha_\circ)} &\leq \|\!| w \|\!|^2_{(A,\rho,\alpha)} + \delta_1\|\nabla w\| + \delta_2\|w\|^2 + \delta_3\|w\|_{\Gamma_R}\\
&\leq \|\!| w \|\!|^2_{(A,\rho,\alpha)} + \frac{\delta_1}{\underline{c}_1-\delta_1}\|\nabla w\|_A^2 + \frac{\delta_2}{\underline{c}_2-\delta_2}\|\sqrt{\rho}w\|^2 + \frac{\delta_3}{\underline{c}_3-\delta_3}\|\sqrt{\alpha}w\|^2_{\Gamma_R}\\
&\leq \max_{i\in\{1,2,3\}}\left(1+\frac{\delta_i}{\underline{c}_i-\delta_i}\right)\ \|\!| w \|\!|^2_{(A,\rho,\alpha)}\ .
\end{aligned}$$

We derive a similar estimate for the lower bound,

$$\begin{aligned}
\|\!| w \|\!|^2_{(A_\circ,\rho_\circ,\alpha_\circ)} &\geq \|\!| w \|\!|^2_{(A,\rho,\alpha)} - \delta_1\|\nabla w\| - \delta_2\|w\|^2 - \delta_3\|w\|_{\Gamma_R}\\
&\geq \|\!| w \|\!|^2_{(A,\rho,\alpha)} + \frac{\delta_1}{\underline{c}_1-\delta_1}\|\nabla w\|_A^2 + \frac{\delta_2}{\underline{c}_2-\delta_2}\|\sqrt{\rho}w\|^2 + \frac{\delta_3}{\underline{c}_3-\delta_3}\|\sqrt{\alpha}w\|^2_{\Gamma_R}\\
&\geq \min_{i\in\{1,2,3\}}\left(1-\frac{\delta_i}{\underline{c}_i-\delta_i}\right)\ \|\!| w \|\!|^2_{(A,\rho,\alpha)}\ .
\end{aligned}$$

An another lower bound is as follows:

$$\begin{aligned}
\|\!| w \|\!|^2_{(A_\circ,\rho_\circ,\alpha_\circ)} &\geq \underline{c}_1\|\nabla w\|^2 + \underline{c}_2\|w\|^2 + \underline{c}_3\|w\|^2_{\Gamma_R}\\
&\geq \frac{\underline{c}_1}{\overline{c}_1+\delta_1}\|\nabla w\|_A^2 + \frac{\underline{c}_2}{\overline{c}_2+\delta_2}\|\sqrt{\rho}w\|^2 + \frac{\underline{c}_3}{\overline{c}_3+\delta_3}\|\sqrt{\alpha}w\|^2_{\Gamma_R}\\
&\geq \min_{i\in\{1,2,3\}}\frac{\underline{c}_i}{\overline{c}_i+\delta_i}\ \|\!| w \|\!|^2_{(A,\rho,\alpha)}\ .
\end{aligned}$$

Clearly, the maximum of the lower bounds is also a lower bound. Definitions of θ_i lead to the statement. □

Theorem 4.3. *Let u be the solution of the reaction diffusion problem (2.23–2.26) and let A, ρ, and α be defined by* (4.30), (4.31) *and* (4.32) *respectively. Then*

$$r^2 \geq \underline{\mathrm{C}}\sup_{w\in V_0}\mathrm{M}^r_\ominus(u_\circ,w), \tag{4.33}$$

where w is an arbitrary function in V_0, $\underline{\mathrm{C}}$ is from Proposition 4.3 and

$$\mathrm{M}^r_\ominus(u_\circ,w) := -\ \|\!| w \|\!|_0^2 + \delta_1\int\limits_\Omega |\nabla w+2\nabla u_\circ|\ |\nabla w|\,dx + \\
+\delta_2\int\limits_\Omega |(w+2u_\circ)w|\,dx + \delta_2\int\limits_{\Gamma_R}|(w+2u_\circ)w|\,\mathrm{ds}. \tag{4.34}$$

Proof. We have

$$r^2 = \sup_{u\in\mathcal{S}} \| u_\circ - u \|_0^2 \geq \underline{C} \sup_{u\in\mathcal{S}} \| u_\circ - u \|_{(A,\rho,\alpha)}^2 . \quad (4.35)$$

On the other hand

$$\begin{aligned} \sup_{u\in\mathcal{S}(\mathcal{D})} \| u_\circ - u \|_{(A,\rho,\alpha)}^2 &= \sup_{(A,\rho,\alpha)\in\mathcal{D}} \left\{ \sup_{w\in V_0} \mathcal{M}_\ominus^{A,\rho,\alpha}(u_\circ, w) \right\} \\ &= \sup_{w\in V_0} \left\{ \sup_{(A,\rho,\alpha)\in\mathcal{D}} \mathcal{M}_\ominus^{A,\rho,\alpha}(u_\circ, w) \right\} \end{aligned}$$

and we conclude that

$$r^2 \geq \underline{C} \sup_{w\in V_0} \left\{ \sup_{(A,\rho,\alpha)\in\mathcal{D}} \mathcal{M}_\ominus^{A,\rho,\alpha}(u_\circ, w) \right\}. \quad (4.36)$$

Now our goal is to estimate the right-hand side of (4.36) from below. For this purpose, we exploit the structure of the minorant, which allows us to explicitly evaluate effects caused by the indeterminacy of the coefficients.

The minorant (from Theorem 2.2) for the reaction diffusion problem can be represented as follows:

$$\begin{aligned} \mathcal{M}_\ominus^{A,\rho,\alpha}(u_\circ, w) = &- \int_\Omega (A_\circ + \delta_1 \Psi)(\nabla w + 2\nabla u_\circ) \cdot \nabla w \, \mathrm{dx} - \\ &- \int_\Omega (\rho_\circ + \delta_2 \psi_\rho)(w + 2u_\circ) w \, \mathrm{dx} - \\ &- \int_{\Gamma_R} (\alpha_\circ + \delta_3 \psi_\alpha)(w + 2u_\circ) w \, \mathrm{ds} + 2l(w), \end{aligned} \quad (4.37)$$

where

$$l(w) := \int_\Omega f w \, \mathrm{dx} - \int_{\Gamma_N} F w \, \mathrm{ds} - \int_{\Gamma_R} G w \, \mathrm{ds}.$$

Note that

$$\int_\Omega (A_\circ \nabla u_\circ \cdot \nabla w \, \mathrm{dx} + \rho_\circ u_\circ w) \, \mathrm{dx} + \int_{\Gamma_R} \alpha_\circ u_\circ w \, \mathrm{ds} = l(w).$$

Hence

$$\begin{aligned} \mathcal{M}_\ominus^{A,\rho,\alpha}(u_\circ, w) = &- \int_\Omega A_\circ \nabla w \cdot \nabla w \, \mathrm{dx} - \int_\Omega \rho_\circ w^2 \, \mathrm{dx} - \int_{\Gamma_R} \alpha_\circ w^2 \, \mathrm{ds} - \\ &- \delta_1 \int_\Omega \Psi(\nabla w + 2\nabla u_\circ) \cdot \nabla w \, \mathrm{dx} - \delta_2 \int_\Omega \psi_\rho (w + 2u_\circ) w \, \mathrm{dx} - \\ &- \delta_3 \int_{\Gamma_R} \psi_\alpha (w + 2u_\circ) w \, \mathrm{ds} \end{aligned} \quad (4.38)$$

and we obtain

$$\sup_{(A,\rho,\alpha)\in\mathcal{D}} \mathcal{M}_{\ominus}^{A,\rho,\alpha}(u_\circ, w) = - \| w \|_\circ^2 + \\ + \delta_1 \sup_{|\Psi|\leq 1} \int\limits_\Omega \Psi(\nabla w + 2\nabla u_\circ) \cdot \nabla w \, \mathrm{dx} + \delta_2 \sup_{|\psi_\rho|\leq 1} \int\limits_\Omega \psi_\rho (w + 2u_\circ) w \, \mathrm{dx} + \\ + \delta_3 \sup_{|\psi_\alpha|\leq 1} \int\limits_{\Gamma_R} \psi_\alpha (w + 2u_\circ) w \, \mathrm{ds}. \quad (4.39)$$

The integrand of the first integral in the right-hand side of (4.39) can be presented as $\Psi : \tau$, where

$$\tau = \nabla w \otimes (\nabla w + 2\nabla u_\circ)$$

and $\otimes$ stands for the diad product. For the first supremum we have

$$\sup_{|\Psi|\leq 1} \left\{ \int\limits_\Omega \Psi : \tau \, \mathrm{dx} \right\} = \int\limits_\Omega |\tau| \, \mathrm{dx}\,. \quad (4.40)$$

Analogously, we find that

$$\sup_{|\psi_\rho|\leq 1} \int\limits_\Omega \psi_\rho (w + 2u_\circ) w \, \mathrm{dx} = \int\limits_\Omega |(w + 2u_\circ) w| \, \mathrm{dx}, \quad (4.41)$$

$$\sup_{|\psi_\alpha|\leq 1} \int\limits_{\Gamma_R} \psi_\alpha (w + 2u_\circ) w \, \mathrm{dx} = \int\limits_{\Gamma_R} |(w + 2u_\circ) w| \, \mathrm{ds}. \quad (4.42)$$

By (4.40)–(4.42), we arrive at the relation

$$\sup_{(A,\rho,\alpha)\in\mathcal{D}} \mathcal{M}_{\ominus}^{A,\rho,\alpha}(u_\circ, w) = - \| w \|_\circ^2 + \delta_1 \int\limits_\Omega |(\nabla w + 2\nabla u_\circ) \otimes \nabla w| \, \mathrm{dx} + \\ + \delta_2 \int\limits_\Omega |(w + 2u_\circ) w| \, \mathrm{dx} + \delta_3 \int\limits_{\Gamma_R} |(w + 2u_\circ) w| \, \mathrm{ds}, \quad (4.43)$$

which together with (4.36) leads to (4.33). □

Theorem 4.3 gives a general form of the lower bound of r. Also, it creates a basis for a practical computation of this quantity. Indeed, let $V_{0h} \subset V_0$ be a finite dimensional space. Then

$$r^2 \geq \underline{C} \sup_{w\in V_{0h}} \mathrm{M}_{\ominus}^{r}(u_\circ, w). \quad (4.44)$$

It is worth noting that the wider set V_{0h} we take, the better lower bound of the radius we compute. If the exact solution $u_\circ$ is not at our disposal, then we can apply an approximation $u_\circ^h \approx u_\circ$. The difference between $u_\circ^h$ and $u_\circ$ enters in the estimate as an additional term that can be estimated by the upper deviation estimate (2.70). This leads to the following Corollary:

Corollary 4.1. *Under the assumptions of Theorem 4.3,*

$$r^2 \geq \mathrm{M}^r_\ominus(u^h_\circ, w) - 2 \interleave u_\circ - u^h_\circ \interleave_\circ \left(\theta_1 \|\nabla w\| + \theta_2 \|\sqrt{\rho} w\| + \theta_3 \|\sqrt{\alpha} w\|_{\Gamma_R}\right)^{\frac{1}{2}}, \forall w \in V_0,$$

where

$$\interleave u_\circ - u^h_\circ \interleave_\circ \leq \sqrt{\mathcal{M}_\oplus(u^h_\circ, y, \gamma, \mu_1, \mu_2)}, \quad \forall y \in Q,\ \gamma > 0,\ \mu_1, \mu_2,$$

where $\mu_i : \Omega \to [0,1]$ *(*$i = \{1,2\}$*).*

Proof. We estimate the lower bound $\mathrm{M}^r_\ominus(u_\circ, w)$ in Theorem 4.1 from below as follows:

$$\begin{aligned} \mathrm{M}^r_\ominus(u_\circ, w) = &- \interleave w \interleave_0^2 + \delta_1 \int_\Omega |\nabla w + 2\nabla u^h_\circ + 2\nabla(u_\circ - u^h_\circ)|\ |\nabla w|\,\mathrm{dx} + \\ &+ \delta_2 \int_\Omega |w + 2u^h_\circ + 2(u_\circ - u^h_\circ)|\ \|w|\,\mathrm{dx} + \delta_2 \int_{\Gamma_R} |(w + 2u^h_\circ + 2(u_\circ - u^h_\circ))w|\,\mathrm{ds} \\ \geq &\ \mathrm{M}^r_\ominus(u^h_\circ, w) - 2\delta_1 \int_\Omega |\nabla(u_\circ - u^h_\circ)|\ |\nabla w|\,\mathrm{dx} - 2\delta_2 \int_\Omega |u_\circ - u^h_\circ|\ |w|\,\mathrm{dx} - \\ &- 2\delta_3 \int_{\Gamma_R} |u_\circ - u^h_\circ|\ |\nabla w|\,\mathrm{ds}. \end{aligned} \tag{4.45}$$

We can estimate each negative term above using bounds for "mean" functions and the Hölder inequality,

$$\begin{aligned} \int_\Omega |\nabla(u_\circ - u^h_\circ)|\ |\nabla w|\,\mathrm{dx} &\leq \frac{1}{\underline{c}_1} \|\nabla(u_\circ - u^h_\circ)\|_{A_\circ} \|\nabla w\| \\ \int_\Omega |u_\circ - u^h_\circ|\ |w|\,\mathrm{dx} &\leq \frac{1}{\underline{c}_2} \|\sqrt{\rho}(u_\circ - u^h_\circ)\|\ \|w\| \\ \int_{\Gamma_R} |u_\circ - u^h_\circ|\ |w|\,\mathrm{ds} &\leq \frac{1}{\underline{c}_3} \|\sqrt{\alpha}(u_\circ - u^h_\circ)\|_{\Gamma_R} \|w\|_{\Gamma_R}. \end{aligned}$$

Substituting these estimates to (4.45) and applying the Cauchy–Schwartz inequality yields the statement. □

A meaningful lower bound can be deduced even analytically as it has been derived for the generalized model earlier in this Section.

Corollary 4.2. *Under assumptions of Theorem 4.3,*

$$r^2 \geq \underline{\mathrm{C}}\, r^2_\ominus \quad \text{and} \quad \hat{r}^2 \geq \underline{\mathrm{C}}\, \hat{r}^2_\ominus, \tag{4.46}$$

where

$$r^2_\ominus = \frac{\interleave u_\circ \interleave_\delta^4}{\interleave u_\circ \interleave_\circ^2 - \interleave u_\circ \interleave_\delta^2} \geq \frac{\Theta^2}{1 - \Theta} \interleave u_\circ \interleave_\circ^2, \tag{4.47}$$

where

$$||| u_\circ |||_\delta^2 := \delta_1 ||\nabla u_\circ||_\Omega^2 + \delta_2 ||u_\circ||_\Omega^2 + \delta_3 ||u_\circ||_{\Gamma_R}^2$$

and

$$\Theta := \min_{i\in\{1,2,3\}} \mathrm{cond}_i\, \theta_i \,.$$

For the normalized radius, we have

$$\hat{r}_\ominus^2 = \frac{\Theta^2}{1-\Theta}. \tag{4.48}$$

Proof. In Theorem 4.3 we estimate supremum over the set V_0 from below by setting

$$w = \lambda u_\circ, \tag{4.49}$$

where $\lambda \in \mathbb{R}$. Then, we observe that

$$r^2 \geq \underline{C}\Bigg(-\lambda^2 ||| u_\circ |||_\circ^2 + \lambda(\lambda+2) ||| u_\circ |||_\delta^2 \Bigg) \,. \tag{4.50}$$

The right–hand side of (4.50) is a quadratic function with respect to λ. It attains its maximal value if

$$\lambda ||| u_\circ |||_\circ^2 = (\lambda+1) ||| u_\circ |||_\delta^2,$$

i.e., if

$$\lambda = \frac{||| u_\circ |||_\delta^2}{||| u_\circ |||_\circ^2 - ||| u_\circ |||_\delta^2}.$$

Substituting this λ, we arrive at (4.47). Note that

$$\begin{aligned} ||| u_\circ |||_\circ^2 &= \int\limits_\Omega (A_\circ \nabla u_\circ \cdot \nabla u_\circ + \rho_\circ u_\circ^2)\,\mathrm{dx} + \int\limits_{\Gamma_R} \alpha_\circ u_\circ \,\mathrm{ds} \\ &\geq \int\limits_\Omega (\underline{c}_1 \nabla u_\circ \cdot \nabla u_\circ + \underline{c}_2 u_\circ^2)\,\mathrm{dx} + \int\limits_{\Gamma_R} \underline{c}_3 u_\circ \,\mathrm{ds} \\ &> \delta_1 ||\nabla u_\circ||_\Omega^2 + \delta_2 ||u_\circ||_\Omega^2 + \delta_3 ||u_\circ||_{\Gamma_R}^2 = ||| u_\circ |||_\delta^2, \end{aligned}$$

so that λ (and the respective lower bound) is positive. Moreover,

$$\begin{aligned} ||| u_\circ |||_\delta^2 \geq \frac{\delta_1}{\bar{c}_1} \int\limits_\Omega A_\circ \nabla u_\circ \cdot \nabla u_\circ \,\mathrm{dx} + \frac{\delta_2}{\bar{c}_2} \int\limits_\Omega \rho_\circ u_\circ^2 \,\mathrm{dx} + \frac{\delta_3}{\bar{c}_3} \int\limits_{\Gamma_R} \alpha_\circ u_\circ^2 \,\mathrm{ds} \\ \geq \Theta ||| u_\circ |||_\circ^2 \,. \end{aligned} \tag{4.51}$$

Also,

$$
\begin{aligned}
&||| \, u_\circ \, |||_\circ^2 - ||| \, u_\circ \, |||_\delta^2 \\
&\quad = \int\limits_\Omega (A_\circ - \delta_1 I)\nabla v \cdot \nabla v \, \mathrm{dx} + \int\limits_\Omega (\rho_\circ - \delta_2) v^2 \, \mathrm{dx} + \int\limits_{\Gamma_R} (\alpha_\circ - \delta_3) v^2 \, \mathrm{ds} \\
&\geq \left(1 - \frac{\delta_1}{\overline{c}_1}\right) \int\limits_\Omega A_\circ \nabla v \cdot \nabla v \, \mathrm{dx} + \left(1 - \frac{\delta_2}{\overline{c}_2}\right) \int\limits_\Omega \rho_\circ v^2 \, \mathrm{dx} + \left(1 - \frac{\delta_3}{\overline{c}_3}\right) \int\limits_{\Gamma_R} \alpha_\circ v^2 \, \mathrm{ds} \\
&\qquad \geq \max_{i=1,2,3} \left(1 - \frac{\delta_i}{\overline{c}_i}\right) ||| \, u_\circ \, |||_\circ^2 = (1 - \Theta) \, ||| \, u_\circ \, |||_\circ^2 \, . \qquad (4.52)
\end{aligned}
$$

By (4.51) and (4.52) we arrive at relation

$$
r_\ominus^2 \geq \frac{\Theta^2}{1 - \Theta} \, ||| \, u_\circ \, |||_\circ^2, \tag{4.53}
$$

which implies (4.47) and (4.48). □

Below, we derive an upper estimate for r, which serves as a natural counterpart for the lower bound derived in Corollary 4.2.

Proposition 4.4. *Let u be the solution of the reaction diffusion problem (2.23–2.26) and let A, ρ, and α be defined by* (4.30), (4.31), *and* (4.32) *respectively. Then*

$$
r^2 \leq \overline{\mathrm{C}} \, r_\oplus^2 \quad \textit{and} \quad \hat{r}^2 \leq \overline{\mathrm{C}} \, \hat{r}_\oplus^2, \tag{4.54}
$$

where

$$
r_\oplus^2 = \frac{\delta_1^2}{\underline{c}_1 - \delta_1} ||\nabla u_\circ||_\Omega^2 + \frac{\delta_2^2}{\underline{c}_2 - \delta_2} ||u_\circ||_\Omega^2 + \frac{\delta_3^2}{\underline{c}_3 - \delta_3} ||u_\circ||_{\Gamma_R}^2 \tag{4.55}
$$

and

$$
\hat{r}_\oplus^2 = \max_{i \in \{1,2,3\}} \frac{\delta_i^2}{\underline{c}_i(\underline{c}_i - \delta_i)}. \tag{4.56}
$$

Proof. By properties of the majorant and (2.7) , we have

$$
\begin{aligned}
\sup_{u \in \mathcal{S}(\mathcal{D})} ||| \, u_\circ - \tilde{u} \, |||_{(A,\rho,\alpha)}^2 &= \sup_{(A,\rho,\alpha) \in \mathcal{D}} \left\{ \inf_{y,\mu_i,\gamma_j} \mathcal{M}_\oplus^{A,\rho,\alpha}(u_\circ, y, \gamma, \mu_1, \mu_2) \right\} \\
&\leq \inf_{y,\mu_i,\gamma_j} \left\{ \sup_{(A,\rho,\alpha) \in \mathcal{D}} \mathcal{M}_\oplus^{A,\rho,\alpha}(u_\circ, y, \gamma, \mu_1, \mu_2) \right\}.
\end{aligned}
$$

Applying Proposition 4.3 we obtain

$$
r^2 \leq \overline{\mathrm{C}} \inf_{y,\mu_i,\gamma_j} \left\{ \sup_{A,\rho,\alpha} \mathcal{M}_\oplus^{A,\rho,\alpha}(u_\circ, y, \gamma, \mu_1, \mu_2) \right\}. \tag{4.57}
$$

Our task is to explicitly estimate the term in the brackets. For this purpose we estimate from above the last two terms of the majorant and represent it in the form

$$\mathcal{M}_{\oplus}^{A,\rho,\alpha}(u_\circ,y,\gamma,\mu_1,\mu_2) \leq$$
$$\kappa\Bigg(\gamma_1 D(\nabla v,y) + \left\|\sqrt{\gamma_2\kappa(1-\mu_1)^2+\frac{\mu_1^2}{\kappa(\underline{c}_2-\delta_2)}}\; r_1(v,y)\right\|_\Omega^2 +$$
$$+\left\|\sqrt{\gamma_3\kappa(1-\mu_2)+\frac{\mu_2}{\kappa(\underline{c}_3-\delta_3)}}\; r_2(v,y)\right\|_{\Gamma_R}^2 + \gamma_4||F-y\cdot\nu\,||_{\Gamma_N}^2\Bigg). \quad (4.58)$$

Now we find upper bounds with respect to $A\in\mathcal{D}_A$, $\rho\in\mathcal{D}_\rho$, $\alpha\in\mathcal{D}_\alpha$ separately.

First, we consider the term D generated by A and A^{-1}:

$$\sup_{A\in\mathcal{D}_A} D(\nabla u_\circ,y) = \sup_{|\Psi|<1}\int_\Omega (A_\circ+\delta_1\Psi)^{-1}|(A_\circ+\delta\Psi)\nabla u_\circ - y|^2\,\mathrm{dx}$$
$$\leq \frac{1}{\underline{c}_1-\delta_1}\sup_{|\Psi|<1}\left\{||A_\circ\nabla u_\circ-y||^2+2\delta_1\int_\Omega \Psi\nabla u_\circ\cdot(A_\circ\nabla u_\circ-y)\,\mathrm{dx}+\delta_1^2||\Psi\nabla u_\circ||^2\right\}$$
$$\leq \frac{1}{\underline{c}_1-\delta_1}\left(||A_\circ\nabla u_\circ-y||_\Omega^2+2\delta_1\int_\Omega|\nabla u_\circ|\,|A_\circ\nabla u_\circ-y|\,\mathrm{dx}+\delta_1^2||\nabla u_\circ||_\Omega^2\right). \quad (4.59)$$

For the term related to the error in the equilibrium equation we have

$$\sup_{\rho\in\mathcal{D}_\rho}||r_1^\rho(u_\circ,y)||_\Omega^2 = \sup_{|\psi_2|<1}\int_\Omega (f-(\rho_\circ+\delta_2\psi_2)u_\circ+\mathrm{div}y)^2\,\mathrm{dx}$$
$$= \sup_{|\psi_2|<1}\int_\Omega \left(\mathrm{div}y-\mathrm{div}(A_\circ\nabla u_\circ)-\delta_2\psi_2 u_\circ\right)^2\,\mathrm{dx}$$
$$\leq ||\mathrm{div}(y-A_\circ\nabla u_\circ)||_\Omega^2+2\delta_2\int_\Omega|\mathrm{div}(y-A_\circ\nabla u_\circ)|\,|u_\circ|\,\mathrm{dx}+\delta_2||u_\circ||^2. \quad (4.60)$$

Similarly, for the term related to the error in the Robin boundary condition we have

$$\sup_{\alpha\in\mathcal{D}_\alpha}||r_2^\alpha(u_\circ,y)||_{\Gamma_R}^2 \leq \left\|\frac{\partial(y-A_\circ\nabla u_\circ)}{\partial\nu}\right\|_{\Gamma_R}^2 +$$
$$+2\delta_3\int_{\Gamma_R}\left|\frac{\partial(y-A_\circ\nabla u_\circ)}{\partial\nu}\right|\,|u_\circ|\,\mathrm{ds}+\delta_3^2||u_\circ||_{\Gamma_R}^2. \quad (4.61)$$

It is clear that for $y=y_0:=A_\circ\nabla u_\circ$ the estimates (4.59)–(4.61) attain minimal values. In addition, we set in (4.58) $\mu_1=\mu_2=1$ and find that

$$\mathcal{M}_{\oplus}^{A,\rho,\alpha}(u_\circ,A_\circ\nabla u_\circ,\gamma,1,1) \leq$$
$$\leq \kappa\left(\frac{\delta_1^2\gamma_1}{\underline{c}_1-\delta_1}||\nabla u_\circ||_\Omega^2+\frac{\delta_2^2}{\underline{c}_2-\delta_2}||u_\circ||_\Omega^2+\frac{\delta_3^2}{\underline{c}_3-\delta_3}||u_\circ||_{\Gamma_R}^2\right). \quad (4.62)$$

Now we tend γ_2, γ_3, and γ_4 (which are contained in κ) to infinity. Then, (4.62) and (4.57) imply (4.55). An upper bound for the normalized radius follows from the relation

$$\begin{aligned} &\mathcal{M}_{\oplus}^{A,\rho,\alpha}(u_\circ, A_\circ\nabla u_\circ, \gamma, 1, 1) \leq \\ &\leq \frac{\delta_1^2}{\underline{c}_1(\underline{c}_1-\delta_1)}\int_\Omega A\nabla u_\circ\cdot\nabla u_\circ\,\mathrm{dx} + \frac{\delta_2^2}{\underline{c}_2(\underline{c}_2-\delta_2)}||\sqrt{\rho_\circ}u_\circ||_\Omega^2 + \frac{\delta_3^2}{\underline{c}_3(\underline{c}3-\delta_3)}||\sqrt{\alpha_\circ}u_\circ||_{\Gamma_R}^2 \\ &\leq \max_{i\in\{1,2,3\}}\frac{\delta_i^2}{\underline{c}_i(\underline{c}_i-\delta_i)} \;|||\; u_\circ \;|||^2, \end{aligned}$$

which leads to (4.56). □

Remark 4.2. *The estimates for the radius of the solution set derived in Corollary 4.2 and Proposition 4.4 for the reaction diffusion problem are very similar to the ones in Theorem 4.1 for the generalized problem. The same similarity is valid for Propositions 4.2 and 4.3. In these estimates, if there are several sources of indeterminacy, the most uncertain parameter dominates.*

4.4 Linear elasticity with incompletely known Poisson's ratio

Here we consider the isotropic linear elasticity problem defined in Section 2.1.2. Our main goal is to estimate the energy sensitivity with respect to ν and to show that for some classes of linear elasticity problems the exact solutions are extremely sensitive to small variations of ν. In this Section, we derive asymptotic estimates that demonstrate a phenomenon that can be called a "blow–up" of the indeterminacy error caused by uncertainty in the material parameters.

Henceforth we assume that the uncertainty of material parameters is generated by one factor, uncertainty of Poisson's ratio ν. In practice, values of E are often known only within some interval, but this constant enters the equation as a multiplier. In view of this fact, the corresponding effects are easy to evaluate (they are proportional to the indeterminacy range). In this study we neglect these effects. Moreover, we assume that solutions are normalized with respect to E, which effectively means that E is replaced by one.

We denote by superscript ν quantities and functions associated with Poisson's ratio ν (e.g., C^ν in Hooke's law (2.15)). Similarly we denote the energy, strain, displacement or stress related to the exact solution of (2.20) (e.g., u^ν denotes the exact solution of (2.20) and ε^ν stands for $\varepsilon(u^\nu)$). We estimate the difference of quantities related to the exact solutions obtained for Poisson's ratios ν and $\nu+\delta$. For this purpose, it is convenient to use incremental relations, e.g.,

$$\triangle_\delta^\nu\mathcal{E} := \frac{\mathcal{E}^{\nu+\delta}-\mathcal{E}^\nu}{\delta} \tag{4.63}$$

for the total energy and the corresponding derivative $\frac{\partial\mathcal{E}}{\partial\nu} = \lim\limits_{\delta\to 0}\triangle_\delta^\nu\mathcal{E}$. The energy $\mathcal{E}$ presents an important integral characteristic of the exact solution. If the solu-

TABLE 12 Analogy of Example 3 and isotropic linear elasticity.

Example 3	Linear elasticity	Physical description
α	ν	Poisson ratio
x	$\epsilon(u)$	strains
y	σ	stresses
K	V_0	set of admissible solutions
Q	$\mathcal{E}$	energy functional
$\frac{1}{1-\alpha}x^T Ex$	$\int_\Omega \lambda \mathrm{div}(u)^2\,\mathrm{dx}$	first part of the energy functional
$\mathcal{N}(E)$	$\mathcal{N}(\mathrm{div})$	kernel of the "blow–up" term
$x^T x$	$\int_\Omega \mu\lvert\epsilon(u)\rvert^2\,\mathrm{dx}$	second part of the energy functional

tion is robust (insensitive) to small variations of the material parameters, then the energy also changes insignificantly. However, if $\triangle^{\nu}_{\delta}\mathcal{E}$ becomes large for relatively small δ, then this fact definitely indicates a high sensitivity of the exact solution to material parameters.

Before advancing to the analysis, recall the Example 3 on page 66, where the relation between the set of admissible solutions and the sensitivity of the energy functional (and the solution) was studied. The analogy between the simple algebraic example and linear elasticity problem is presented in Table 12.

The function $Q : \mathbb{R}^2 \to \mathbb{R}$ is an analog of the energy functional $\mathcal{E}$ in the linear elasticity theory. The term $\frac{1}{\alpha}x^T Ex$ behaves like the divergence term $\int_\Omega \lambda \mathrm{div}(u)^2\,\mathrm{dx}$ in two ways: The coefficient tends to infinity as parameter α tends to 1 in a similar way as λ tends to infinity as ν tends to $\frac{1}{2}$ (the incompressibility limit). The kernel is not trivial; in other words there are non-zero vectors (functions in linear elasticity case), for which the value of this term is zero. The term $x^T x$ is zero only if $x = 0$ and the coefficient is bounded. It behaves like the strain term $\int_\Omega \mu\lvert\epsilon(u)\rvert^2\,\mathrm{dx}$ in the space $H^1_0(\Omega)$, which is zero only if $u = 0$.

The two cases $K \cap \mathcal{N}(E) \neq \emptyset$ and $K \cap \mathcal{N}(E) = \emptyset$ are analogs of two different types of boundary conditions: The first one admits the existence of divergence–free solutions and the second one does not. In linear elasticity the analytical solutions are not usually at our disposal. For this reason the analysis exposed in this Section is much more complicated.

Estimates of the sensitivity of energy to Poisson's ratio

Henceforth, we assume that

$$0 \le \nu < \tfrac{1}{2} \quad \text{and} \quad \delta \le \tfrac{1}{2}\left(\tfrac{1}{2} - \nu\right). \tag{4.64}$$

This condition guarantees that the problems with different Poisson's coefficients are uniformly elliptic. Also, we assume $\ell = 0$ (this assumption is made only for the sake of simplicity).

First, we establish the relation that serves as a basis for our subsequent analysis. We have

$$\begin{aligned}\interleave \varepsilon^{\nu+\delta} - \varepsilon^{\nu} \interleave^2_{C^{\nu+\delta}} &= \int\limits_\Omega C^{\nu+\delta}\varepsilon^{\nu} : \varepsilon^{\nu}\, \mathrm{dx} - \int\limits_\Omega C^{\nu+\delta}\varepsilon^{\nu+\delta} : \varepsilon^{\nu+\delta}\, \mathrm{dx} \\ &= \delta \int\limits_\Omega (\triangle^{\nu}_{\delta} C)\varepsilon^{\nu} : \varepsilon^{\nu}\, \mathrm{dx} - \delta(\triangle^{\nu}_{\delta}\mathcal{E}).\end{aligned}$$

Hence

$$\triangle^{\nu}_{\delta}\mathcal{E} = \int\limits_\Omega (\triangle^{\nu}_{\delta} C)\varepsilon^{\nu} : \varepsilon^{\nu}\, dx - \frac{1}{\delta} \interleave \varepsilon^{\nu+\delta} - \varepsilon^{\nu} \interleave^2_{C^{\nu+\delta}} . \tag{4.65}$$

For isotropic media we have

$$\int\limits_\Omega (\triangle^{\nu}_{\delta} C)\varepsilon^{\nu} : \varepsilon^{\nu}\, dx = \int\limits_\Omega \Big((\triangle^{\nu}_{\delta}\lambda)|\mathrm{div} u^{\nu}|^2 + 2(\triangle^{\nu}_{\delta}\mu)|\varepsilon^{\nu}|^2 \Big)\, \mathrm{dx}.$$

In view of (4.64) we have

$$\begin{aligned}\kappa_{\ominus} := \frac{4(1+2\nu^2)}{(5+2\nu)(1+\nu)(1-2\nu)^2} \leq \triangle^{\nu}_{\delta}\lambda &\leq \frac{2(1+2\nu^2+2\nu\delta)}{(1+\nu)^2(1-2\nu)^2} \\ &\leq \frac{2+v+2\nu^2}{(1+\nu)^2(1-2\nu)^2} := \kappa_{\oplus}\end{aligned}$$

and

$$m_{\ominus} := -\frac{1}{(1+\nu)^2} \leq 2\triangle^{\nu}_{\delta}\mu \leq -\frac{4}{(5+2\nu)(1+\nu)} := m_{\oplus}.$$

Now we find that

$$\int\limits_\Omega (\triangle^{\nu}_{\delta} C)\varepsilon^{\nu} : \varepsilon^{\nu}\, dx \geq \int\limits_\Omega \left((\kappa_{\ominus}|\mathrm{div} u^{\nu}|^2 - \frac{1}{(1+\nu)^2}|\varepsilon^{\nu}|^2 \right)\, \mathrm{dx}. \tag{4.66}$$

In order to estimate the second term in the right–hand side of (4.65), we use the majorant for the linear elasticity problem,

$$(\mathcal{M}_{\oplus}(v,\tau))^{\frac{1}{2}} = \left(\int\limits_\Omega (C^{\nu+\delta}\varepsilon(u^{\nu}) - \tau) : (\varepsilon(u^{\nu}) - (C^{\nu+\delta})^{-1}\tau)\, \mathrm{dx} \right)^{\frac{1}{2}} + c_K \|\mathrm{Div}\tau + f\|,$$

where c_K is a constant related to Korn's inequality and τ is an auxiliary stress function that is at our disposal. Here, we consider u^{ν} as an approximation v for the problem defined by $u^{\nu+\delta}$. For $\tau := C^{\nu}\varepsilon(u^{v})$, the equilibrium term vanishes and the estimate reads as follows:

$$\interleave \varepsilon(u^{\nu+\delta} - u^{\nu}) \interleave^2_{C^{\nu+\delta}} \leq \mathcal{M}^{\nu+\delta}_{\oplus}(u^{\nu}, \tau), \tag{4.67}$$

where

$$\mathcal{M}^{\nu+\delta}_{\oplus}(u^{\nu}, \tau) := \int\limits_\Omega (C^{\nu+\delta}\varepsilon(u^{\nu}) - \tau) : (\varepsilon(u^{\nu}) - (C^{\nu+\delta})^{-1}\tau)\, \mathrm{dx}.$$

Lemma 4.1.

$$\interleave\, \varepsilon^{\nu+\delta} - \varepsilon^{\nu} \,\interleave^2_{C^{\nu+\delta}} \leq \delta^2 \left(c_1^{\nu} \|\mathrm{div} u^{\nu}\|^2 + c_2^{\nu} \|\varepsilon(u^{\nu})\|^2 \right),$$

where

$$c_1^{\nu} := \frac{\lambda^{\nu}}{\nu}\, n\, \kappa_{\oplus} \quad \text{and} \quad c_2^{\nu} := \tfrac{1}{4}(5+2\nu) m_{\ominus}^2.$$

Proof. It is easy to see that

$$C^{\nu+\delta}\varepsilon(u^{\nu}) - C^{\nu}\varepsilon(u^{\nu}) = (\lambda^{\nu+\delta} - \lambda^{\nu})\mathrm{div} u^{\nu}\mathbb{I} + 2(\mu^{\nu+\delta} - \mu^{\nu})\varepsilon(u^{\nu}). \tag{4.68}$$

Since

$$C^{-1}\tau = \frac{1}{2\mu}\left(\tau - \frac{\lambda}{3\lambda + 2\mu}\mathrm{tr}(\tau)\mathbb{I} \right)$$

we have

$$\begin{aligned}
(C^{\nu+\delta})^{-1}C^{\nu}\varepsilon(u^{\nu}) &= \\
&= \frac{1}{2\mu^{\nu+\delta}}\left(\lambda^{\nu}\mathrm{div} u^{\nu}\mathbb{I} + 2\mu^{\nu}\epsilon(u^{\nu}) - \frac{\lambda^{\nu+\delta}}{3\lambda^{\nu+\delta} + 2\mu^{\nu+\delta}}(3\lambda^{\nu} + 2\mu^{\nu})\mathrm{div} u^{\nu}\mathbb{I} \right) \\
&= \left(\frac{\lambda^{\nu}}{2\mu^{\nu+\delta}} - \frac{\lambda^{\nu+\delta}(3\lambda^{\nu} + 2\mu^{\nu})}{2\mu^{\nu+\delta}(3\lambda^{\nu+\delta} + 2\mu^{\nu+\delta})} \right)\mathrm{div} u^{\nu}\mathbb{I} + \frac{\mu^{\nu}}{\mu^{\nu+\delta}}\,\varepsilon(u^{\nu}).
\end{aligned}$$

We note that

$$3\lambda^{\nu} + 2\mu^{\nu} = \frac{1}{1-2\nu}, \quad \frac{\lambda^{\nu+\delta}}{2\mu^{\nu+\delta}} = \frac{\nu+\delta}{1-2\nu-2\delta},$$

$$\frac{\lambda^{\nu}}{2\mu^{\nu+\delta}} - \frac{\lambda^{\nu+\delta}(3\lambda^{\nu} + 2\mu^{\nu})}{2\mu^{\nu+\delta}(3\lambda^{\nu+\delta} + 2\mu^{\nu+\delta})} = -\frac{\delta}{(1+\nu)(1-2\nu)} =: -\frac{\delta}{\nu}\,\lambda^{\nu},$$

and

$$\varepsilon(u^{\nu}) - (C^{\nu+\delta})^{-1}C^{\nu}\varepsilon(u^{\nu}) = \delta\frac{\lambda^{\nu}}{\nu}\,\mathrm{div} u^{\nu}\mathbb{I} + \frac{\mu^{\nu+\delta} - \mu^{\delta}}{\mu^{\nu+\delta}}\,\varepsilon(u^{\nu}). \tag{4.69}$$

By (4.68) and (4.69) we obtain

$$\begin{aligned}
\mathcal{M}_{\oplus}^{\nu+\delta}(u^{\nu}, \tau) &= \\
&= n\delta\frac{\lambda^{\nu}}{\nu}(\lambda^{\nu+\delta} - \lambda^{\nu})\|\mathrm{div} u^{\nu}\|^2 + \frac{(\lambda^{\nu+\delta} - \lambda^{\nu})(\mu^{\nu+\delta} - \mu^{\nu})}{\mu^{\nu+\delta}}\|\mathrm{div} u^{\nu}\|^2 + \\
&\quad + 2\delta\frac{\lambda^{\nu}}{\nu}(\mu^{\nu+\delta} - \mu^{\nu})\|\mathrm{div} u^{\nu}\|^2 + 2\frac{(\mu^{\nu+\delta} - \mu^{\nu})^2}{\mu^{\nu+\delta}}\|\varepsilon(u^{\nu})\|^2.
\end{aligned}$$

Since the second and third terms are negative, we find that

$$\begin{aligned}
\mathcal{M}_{\oplus}^{\nu+\delta}(u^{\nu}, \tau) &\leq n\delta\frac{\lambda^{\nu}}{\nu}(\lambda^{\nu+\delta} - \lambda^{\nu})\|\mathrm{div} u^{\nu}\|^2 + 2\frac{(\mu^{\nu+\delta} - \mu^{\nu})^2}{\mu^{\nu+\delta}}\|\varepsilon(u^{\nu})\|^2 \\
&\leq \delta^2\left(\frac{\lambda^{\nu}}{\nu}n\kappa_{\oplus}\|\mathrm{div} u^{\nu}\|^2 + \frac{m_{\ominus}^2}{2\mu^{\nu+\delta}}\|\varepsilon(u^{\nu})\|^2 \right),
\end{aligned}$$

which together with (4.67) leads to the statement. □

Theorem 4.4. *Let $u_g \in V$ be a function with the minimal divergence norm, i.e.,*

$$\|\mathrm{div} u_g\| = \min_{u \in V} \|\mathrm{div} u\|. \tag{4.70}$$

Then for $0 \leq \delta \leq \min\{\frac{1}{2c_1^\nu}\kappa_\ominus, \bar{\delta}\}$, the following estimate is valid:

$$\triangle_\delta^\nu \mathcal{E} \geq \frac{1}{2}\kappa_\ominus \|\mathrm{div} u_g\|^2 - c_3^\nu \|\varepsilon(u_g)\|^2, \tag{4.71}$$

where

$$\begin{aligned} c_3^\nu &:= \frac{1}{(1+\nu)^2} + \frac{m_\ominus^2(5+2\nu)}{4}\frac{\nu\kappa_\ominus}{2\lambda^\nu n\kappa_\oplus} \\ &= \frac{1}{(1+\nu)^2}\left(1 + \frac{(1+2\nu^2)(1-2\nu)}{2n(2+\nu+2\nu^2)}\right). \end{aligned} \tag{4.72}$$

Proof. Applying estimate (4.66) and Lemma 4.1 to (4.65) yields

$$\begin{aligned} \triangle_\delta^\nu \mathcal{E} &= \int_\Omega (\triangle_\delta^\nu C)\varepsilon^\nu : \varepsilon^\nu \, dx - \frac{1}{\delta} \mid\mid\mid \varepsilon^{\nu+\delta} - \varepsilon^\nu \mid\mid\mid^2_{C^{\nu+\delta}} \\ &\geq (\kappa_\ominus - \delta c_1^\nu)\|\mathrm{div} u^\nu\|^2 - \left(\frac{1}{(1+\nu)^2} + c_2^\nu \delta\right)\|\varepsilon(u^\nu)\|^2. \end{aligned}$$

Let $\delta \leq \frac{\kappa_\ominus}{2c_1^\nu}$. Then we obtain

$$\begin{aligned} \triangle_\delta^\nu \mathcal{E} &\geq \frac{1}{2}\kappa_\ominus \|\mathrm{div} u^\nu\|^2 - \left(\frac{1}{(1+\nu)^2} + \frac{m_\ominus^2(5+2\nu)}{4}\frac{\nu\kappa_\ominus}{2\lambda^\nu n\kappa_\oplus}\right)\|\varepsilon(u^\nu)\|^2 \\ &= \frac{1}{2}\kappa_\ominus \|\mathrm{div} u^\nu\|^2 - c_3^\nu \|\varepsilon(u^\nu)\|^2. \end{aligned} \tag{4.73}$$

Since u_g satisfies (4.70) and u^ν minimizes the energy functional $\mathcal{E}^\nu$, we find that

$$\begin{aligned} \int_\Omega \left(\lambda^\nu |\mathrm{div}(u_g)|^2 + 2\mu^\nu |\varepsilon(u^\nu)|^2\right) dx &\leq \mathcal{E}^\nu(u^\nu) \leq \mathcal{E}^\nu(u_g) \\ &= \int_\Omega \left(\lambda^\nu |\mathrm{div}(u_g)|^2 + 2\mu^\nu |\varepsilon(u_g)|^2\right) dx \end{aligned} \tag{4.74}$$

and, therefore,

$$\|\varepsilon(u^\nu)\| \leq \|\varepsilon(u_g)\|. \tag{4.75}$$

Applying (4.70) and (4.75), we estimate the r.h.s. of (4.73) from below and arrive at the statement. □

Corollary 4.3. *Since the right–hand side does not depend on δ, we can pass to the limit as $\delta \to 0$ and obtain*

$$\frac{\partial \mathcal{E}^\nu}{\partial \nu} \geq \frac{1}{2}\kappa_\ominus \|\mathrm{div} u_g\|^2 - c_3^\nu \|\varepsilon(u_g)\|^2. \tag{4.76}$$

Theorem 4.4 shows *how sensitive is the internal energy associated with the exact solution* u^ν *with respect to small variations of Poisson's ratio* ν. If it is large, then ν must be known with a high accuracy; otherwise a quantitative analysis of the problem is not motivated. For this reason the estimate (4.76) deserves special discussion. First of all, we note that the right–hand side of (4.76) is easy to compute.

The asymptotic properties of the estimate depend crucially on the first term. A distinctive feature of the set V (of admissible displacements) is whether the boundary conditions are "compatible" (in the sense that there exists a divergence–free function u_g that satisfies these conditions) or not.

If $\mathrm{div} u_g = 0$, then (4.74) shows that

$$\mathcal{E}^\nu(u^\nu) \leq \mathcal{E}^\nu(u_g) = \frac{1}{1+\nu}\|\varepsilon(u_g)\|^2,$$

which means that for all $\nu \in [0, \frac{1}{2})$ such quantities as $\mathcal{E}^\nu$ and $\|\varepsilon(u^\nu)\|$ are uniformly bounded. Moreover,

$$\int\limits_\Omega |\mathrm{div}(u^\nu)|^2 dx \leq \int\limits_\Omega \left(2\frac{\mu^\nu}{\lambda^\nu}|\varepsilon(u_g)|^2\right) dx \to 0 \quad \text{as } \nu \to \frac{1}{2}.$$

In this case, small variations of Poisson's ratio do not lead to large changes in the solution.

Let us consider another case. Assume that the boundary conditions are non-compatible, i.e.,

$$\|\mathrm{div}\, u_g\| > 0. \tag{4.77}$$

Then, even the normalized energy increment blows up.

Corollary 4.4. *Under the assumptions of Theorem 4.4,*

$$\frac{\triangle^\nu_\delta \mathcal{E}^\nu}{\mathcal{E}^\nu} \geq \mathbb{C}(\nu, u_g) \quad \text{and} \quad \frac{1}{\mathcal{E}^\nu}\frac{\partial \mathcal{E}^\nu}{\partial \nu} \geq \mathbb{C}(\nu, u_g),$$

where

$$\mathbb{C}(\nu, u_g) := (1+\nu)\frac{c_4^\nu\|\mathrm{div} u_g\|^2 - c_3^\nu(1-2\nu)\|\varepsilon(u_g)\|^2}{\nu\|\mathrm{div} u_g\|^2 + (1-2\nu)\|\varepsilon(u_g)\|^2}$$

and

$$c_4^\nu := \frac{2(1+2\nu^2)}{(5+2\nu)(1+\nu)(1-2\nu)}. \tag{4.78}$$

Proof. By (4.71) and (4.74) we see that the normalized energy increment is subject to the relation

$$\frac{\triangle^\nu_\delta \mathcal{E}^\nu}{\mathcal{E}^\nu} \geq \frac{\frac{1}{2}\kappa_\ominus\|\mathrm{div} u_g\|^2 - c_3^\nu\|\varepsilon(u_g)\|^2}{\lambda^\nu\|\mathrm{div}(u_g)\|^2 + 2\mu^\nu\|\varepsilon(u_g)\|^2} \tag{4.79}$$

$$= (1+\nu)\frac{c_4^\nu\|\mathrm{div} u_g\|^2 - c_3^\nu(1-2\nu)\|\varepsilon(u_g)\|^2}{\nu\|\mathrm{div} u_g\|^2 + (1-2\nu)\|\varepsilon(u_g)\|^2}. \tag{4.80}$$

Since the right–hand side of (4.79) does not depend on δ, it also follows that the logarithmic derivative of the energy is bounded by the same constant. □

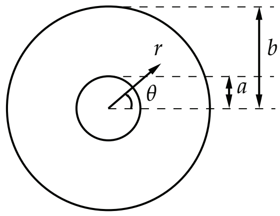

FIGURE 12 A model axisymmetric problem.

$\mathbb{C}(\nu, u_g)$ is the sensitivity constant that depends only on the geometry, ν, and the "minimal divergence function" u_g. If $\|\text{div} u_g\| > 0$, then $\mathbb{C}(\nu, u_g)$ blows up as ν tends to $1/2$. Thus, the value of the energy increment normalized by the value of $\mathcal{E}^\nu$ (and the logarithmic derivative) also becomes highly sensitive to small changes in Poisson's ratio.

Remark 4.3. *It is worth noting that the estimate (4.79) and the constant $\mathbb{C}(\nu, u_g)$ can be used to estimate the effects caused by variations of ν around some given value. Indeed,*

$$\frac{\mathcal{E}^{\nu+\delta} - \mathcal{E}^\nu}{\mathcal{E}^\nu} \geq \delta\, \mathbb{C}(\nu, u_g). \tag{4.81}$$

From (4.81) it follows that if γ denotes the upper limit of acceptable uncertainty in terms of the energy , then the value of Poisson's ratio must be known with the accuracy $\gamma\, \mathbb{C}^{-1}(\nu, u_g)$. Obviously, in the case of a blow up situation this condition may be impossible to satisfy in practice.

4.4.1 Axisymmetric model

In this section, we study exact solutions of an axisymmetric problem and use them to demonstrate effects caused by an incomplete knowledge of Poisson's ratio. The geometry of the problem is presented in Figure 12. Let (r, θ) be in polar coordinates; then

$$\Omega := \{a \leq r \leq b,\ 0 \leq \theta < 2\pi\}.$$

In the polar coordinate system

$$\varepsilon_r = \frac{du_r}{dr}, \quad \varepsilon_\theta = \frac{u_r}{r}, \quad \varepsilon_{r\theta} = \frac{r}{2}\frac{d}{dr}\left(\frac{u_\theta}{r}\right), \tag{4.82}$$

and the constitutive relations read as follows:

$$\sigma_r = 2\mu\varepsilon_r + \lambda(\varepsilon_r + \varepsilon_\theta), \tag{4.83}$$
$$\sigma_\theta = 2\mu\varepsilon_\theta + \lambda(\varepsilon_r + \varepsilon_\theta), \tag{4.84}$$
$$\sigma_{r\theta} = 2\mu\varepsilon_{r\theta}. \tag{4.85}$$

We assume that the volume and surface loads are zero. In this case the equations of equilibrium have the form

$$\frac{d\sigma_r}{dr} + \frac{\sigma_r - \sigma_\theta}{r} = 0, \tag{4.86}$$

$$\frac{d\sigma_{r\theta}}{dr} + 2\frac{\sigma_{r\theta}}{r} = 0, \tag{4.87}$$

and

$$\mathcal{E} := \int_\Omega \sigma : \varepsilon \,\mathrm{dx} = \int_\Omega \left(\lambda \mathrm{tr}^2(\varepsilon) + 2\mu|\varepsilon|^2\right) \mathrm{dx}.$$

First, note that

$$\sigma_r - \sigma_\theta = 2\mu(\varepsilon_r - \varepsilon_\theta) = 2\mu\left(\frac{du_r}{dr} - \frac{u_r}{r}\right) = 2\mu r \frac{d}{dr}\left(\frac{u_r}{r}\right).$$

Substituting this relation to (4.86) we obtain

$$\sigma_r + 2\mu\frac{u_r}{r} = \gamma_1, \tag{4.88}$$

where γ_1 is constant. In view of (4.83), we arrive at the differential equation

$$\frac{d}{dr}(ru_r) = r\,\frac{\gamma_1}{\lambda + 2\mu}$$

which implies

$$u_r = \gamma_1 r + \frac{\gamma_2}{r}. \tag{4.89}$$

From (4.87) we find that $\sigma_{r\theta} = \frac{\gamma_3}{r^2}$. and $u_\theta = \gamma_3 + \frac{\gamma_4}{r}$. Next,

$$\sigma_r = 2(\mu + \lambda)\gamma_1 - 2\mu\frac{\gamma_2}{r^2} \quad \text{and} \quad \sigma_\theta = 2(\mu + \lambda)\gamma_1 + 2\mu\frac{\gamma_2}{r^2}.$$

The energy is

$$\mathcal{E} = 4\pi\left((\lambda + \mu)(b^2 - a^2)\gamma_1^2 + \mu\left(\frac{1}{a^2} - \frac{1}{b^2}\right)\gamma_2^2\right).$$

Compatibility of the set V

As we have discussed earlier, the boundary conditions distinguish whether the set of admissible functions V has (is "compatible") or does not have (is "non–compatible") any divergence–free members. The divergence–free functions for the model problem can be directly computed. They all have the form

$$w(r) := \frac{c}{r},$$

where c is constant.

For the Dirichlet–Neumann boundary conditions,

$$u_r(a) = g_a \quad \text{and} \quad \sigma_r(b) = F_b,$$

the divergence–free minimizer

$$u_g^{DN} = \frac{g_a a}{r}$$

always belongs to the set V. The exact solution of the problem is

$$u^{DN} = \gamma_1^{DN} r + \frac{\gamma_2^{DN}}{r},$$

where

$$\gamma_1^{DN} = \frac{ab^2\,(2g_a(\lambda+\mu) - bF_b)}{2(b^2(\lambda+\mu)+a^2\mu)} \quad \text{and} \quad \gamma_2^{DN} = \frac{F_b b^2 + 2g_a a\mu}{2(b^2(\lambda+\mu)+a^2\mu)}.$$

For the Dirichlet-Dirichlet conditions,

$$u_r(a) = g_a \quad \text{and} \quad u_r(b) = g_b,$$

the divergence minimizer u_g coincides with the solution of the problem and is

$$u^{DD} = u_g^{\mathrm{DD}} = \gamma_1^{DD} r + \frac{\gamma_2^{DD}}{r},$$

where

$$\gamma_1^{DD} = \frac{g_b b - g_a a}{b^2 - a^2} \quad \text{and} \quad \gamma_2^{DD} = \frac{ab\,(g_a b - g_b a)}{b^2 - a^2}.$$

The function u_g is divergence–free only if $g_a a = g_b b$. This condition defines whether the Dirichlet-Dirichlet conditions are "compatible" or not.

Next, we observe the blow–up that occurs with "non–compatible" boundary conditions. We study the behavior of the energy quotient at the incompressibility limit. For our model problem we can compute the derivative of energy with respect to Poisson's ratio,

$$\begin{aligned}\frac{\partial \mathcal{E}^\nu}{\partial \nu} &= 4\pi \left\{ (b^2 - a^2)\left(\frac{\partial \lambda^\nu}{\partial \nu} + \frac{\mu^\nu}{\partial \nu}\right)\gamma_1^2 + \left(\frac{1}{a^2} - \frac{1}{b^2}\right)\frac{\partial \mu^\nu}{\partial \nu}\gamma_2^2 \right\} \\ &= \frac{2\pi(b^2-a^2)}{(1+\nu^2)} \left\{ \frac{1+4\nu}{(1-2\nu)^2}\gamma_1^2 - \frac{1}{a^2b^2}\gamma_2^2 \right\}.\end{aligned}$$

Moreover, for the logarithmic derivative we have

$$\frac{1}{\mathcal{E}^\nu}\frac{\partial \mathcal{E}^\nu}{\partial \nu} = \frac{1}{(1+\nu)(1-2\nu)} \frac{(1+4\nu)a^2b^2\gamma_1^2 - (1-2\nu)^2\gamma_2^2}{a^2b^2\gamma_1^2 + (1-2\nu)\gamma_2^2}.$$

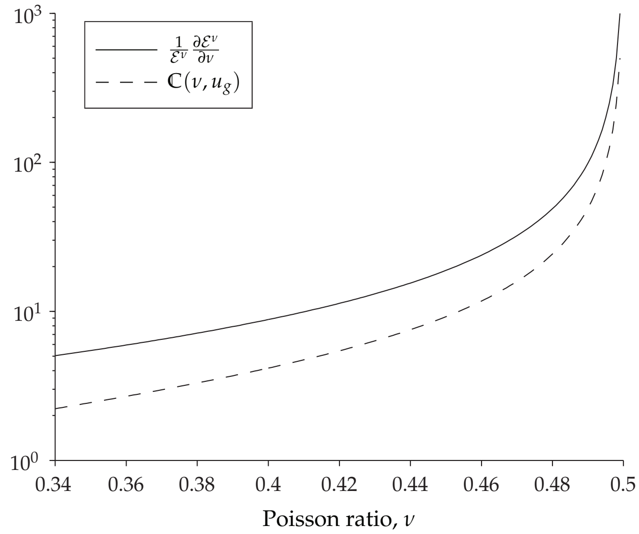

FIGURE 13 Exact values of $\frac{1}{\mathcal{E}}\frac{\partial \mathcal{E}^\nu}{\partial \nu}$ and the lower estimate of Corollary 4.4 for the model problem with pure Dirichlet conditions and parameter values $a = 0.2$, $b = 1.0$, $g_a = 0.01$, and $g_b = -0.03$.

It is easy to see that $\left|\frac{\partial \mathcal{E}^\nu}{\partial \nu}\right|$ and $\left|\frac{1}{\mathcal{E}^\nu}\frac{\partial \mathcal{E}^\nu}{\partial \nu}\right|$ tend to ∞ if $\gamma_1^2 > 0$, i.e., the boundary conditions are non–compatible. If the boundary conditions are compatible, i.e., $g_a a = g_b b$ and $\gamma_1 = 0$, then

$$\left|\frac{1}{\mathcal{E}^\nu}\frac{\partial \mathcal{E}^\nu}{\partial \nu}\right| = \frac{1}{1+\nu}.$$

Moreover, since u_g is known for the applied boundary conditions, we can compute the sensitivity constant in Corollary 4.4,

$$\mathbb{C}(\nu, u_g) = (1+\nu)\frac{a^2b^2\left(2c_4^\nu - (1-2\nu)c_3^\nu\right)\gamma_1 - c_3^\nu(1-2\nu)\gamma_2}{a^2b^2\gamma_1^2 + (1-2\nu)\gamma_2},$$

where c_3^ν and c_4^ν are defined by (4.72) and (4.78). For pure Dirichlet conditions, the blow–up of the exact logarithmic derivative and the sensitivity constant can be observed from Figure 13. For Dirichlet–Neumann conditions, a similar plot provides an exact quotient that is almost zero while the lower bound crudely underestimates it.

Numerically constructed solution set

Here we demonstrate how the sensitivity of the solution depends on the compatibility of the boundary conditions. We do not restrict this study only to demonstrate the blow–up phenomenon. In particular, it is interesting to observe how far from the incompressibility limit the sensitivity of the solution renders any quantitative results too inaccurate for engineering purposes.

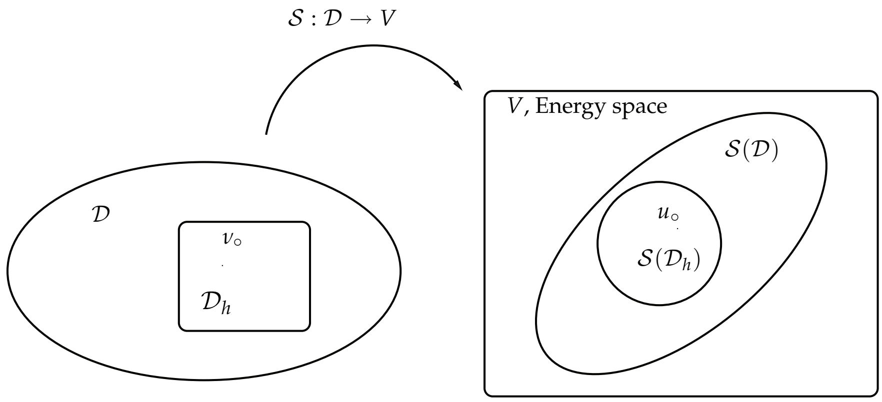

FIGURE 14 Illustration of the experiment.

In reality, ν is an unknown function from the interval (a, b) to some subset of interval $(0, \frac{1}{2})$. Motivated by engineering practices we define this subset by a "mean value" $\nu_\circ$ and limited error bounds. Thus ν belongs to the following family of functions:

$$\nu \in \mathcal{D} := \{\nu : (a, b) \to [\nu_\circ - \delta, \nu_\circ + \delta]\}.$$

For every $\nu \in \mathcal{D}$ there exists a solution. We denote the solution mapping by

$$\mathcal{S} : \mathcal{D} \to V.$$

Our interest is to study the set of solutions $\mathcal{S}(\mathcal{D})$ associated with all members of $\mathcal{D}$. Obviously it is impossible to derive analytical solutions for arbitrary $\nu \in \mathcal{D}$. In order to obtain a reasonable representation of $\mathcal{S}(\mathcal{D})$ we consider the set of piecewise constant functions $\mathcal{D}_h \subset \mathcal{D}$. For $\nu \in \mathcal{D}_h$, we can compute exact solutions and obtain $\mathcal{S}(\mathcal{D}_h)$ The sets are depicted in Figure 14. The procedure for computing these solutions is described below.

Consider the case where the material parameters on the entire domain (a, b) are piecewise constants. Let the interval (a, b) be divided into N non-intersecting subintervals $I_k := (r_k, r_{k+1})$, where r_k ($k = 1, \ldots, N + 1$) are grid points. We assume that Poisson's ratio ν_k is constant on each interval I_k. Moreover, we allow only M different constants from the interval $[\nu_\circ - \delta, \nu_\circ + \delta]$. Consequently, Lame's parameters λ_k and μ_k are piecewise constants too.

Without body forces, on each interval I_k the displacement and stresses have the form:

$$u_r = \gamma_1^k r + \frac{\gamma_2^k}{r},$$
$$\sigma_r = 2(\lambda_k + \mu_k)\gamma_1^k - 2\mu_k \frac{\gamma_2^k}{r^2}.$$

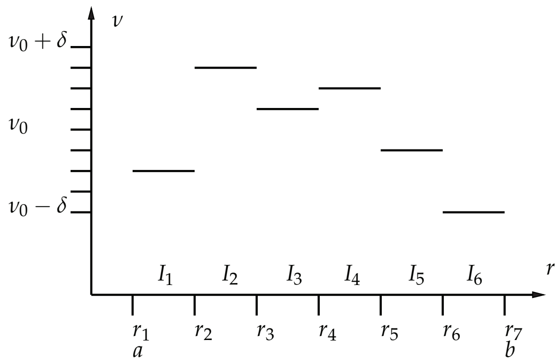

FIGURE 15 Division to subintervals with $N = 6$ and $M = 9$, and an example of the distribution $\nu \in \mathcal{D}_h$.

The solution must satisfy continuity conditions at the junctions of each interval. Thus, for every r_k, $(k = 2, \ldots, N+1)$

$$\gamma_1^{k-1} r_k + \frac{\gamma_2^{k-1}}{r_k} = \gamma_1^k r_k + \frac{\gamma_2^k}{r_k}, \tag{4.90}$$

$$2(\lambda_{k-1} + \mu_{k-1})\gamma_1^{k-1} - 2\mu_{k-1}\frac{\gamma_2^{k-1}}{r_k^2} = 2(\lambda_k + \mu_k)\gamma_1^k - 2\mu_k \frac{\gamma_2^k}{r_k^2}. \tag{4.91}$$

The boundary conditions and continuity conditions (4.90) and (4.91) together form a set of linear equations, from which coefficients γ_1^k and γ_2^k can be solved.

The experiments are performed as follows: We compute a solution associated with all combinations generated by a piecewise continuous $\nu \in \mathcal{D}_h$. A particular member of $\mathcal{D}_h$ is depicted in Figure 15. Each of these solutions (displacement and stresses denoted by subscript pert), are then compared to the non-perturbed solution. We examine the relative perturbations of various solution–dependent quantities defined as follows:

$$e^u_{L_2} := \max_{u_{\text{pert}} \in \mathcal{S}(\mathcal{D}_h)} \frac{\|u_{r0} - u_{\text{pert}}\|_{L_2}}{\|u_{r0}\|_{L_2}}, \qquad e^u_{L_\infty} := \max_{u_{\text{pert}} \in \mathcal{S}(\mathcal{D}_h)} \frac{\|u_{r0} - u_{\text{pert}}\|_{L_\infty}}{\|u_{r0}\|_{L_\infty}},$$

$$e^\sigma_{L_2} := \max_{u_{\text{pert}} \in \mathcal{S}(\mathcal{D}_h)} \frac{\|\sigma_{r0} - \sigma_{\text{pert}}\|_{L_2}}{\|\sigma_{r0}\|_{L_2}}, \qquad e^\sigma_{L_\infty} := \max_{u_{\text{pert}} \in \mathcal{S}(\mathcal{D}_h)} \frac{\|\sigma_{r0} - \sigma_{\text{pert}}\|_{L_\infty}}{\|\sigma_{r0}\|_{L_\infty}}.$$

Bounds of the set $\mathcal{S}(\mathcal{D}_h)$ are also of special interest,

$$u_r^{\min}(r) := \min_{u_{\text{pert}} \in \mathcal{S}(\mathcal{D}_h)} u_{\text{pert}}(r)$$

$$u_r^{\max}(r) := \max_{u_{\text{pert}} \in \mathcal{S}(\mathcal{D}_h)} u_{\text{pert}}(r),$$

and similarly for stress. Moreover, we compute the relative perturbations in the

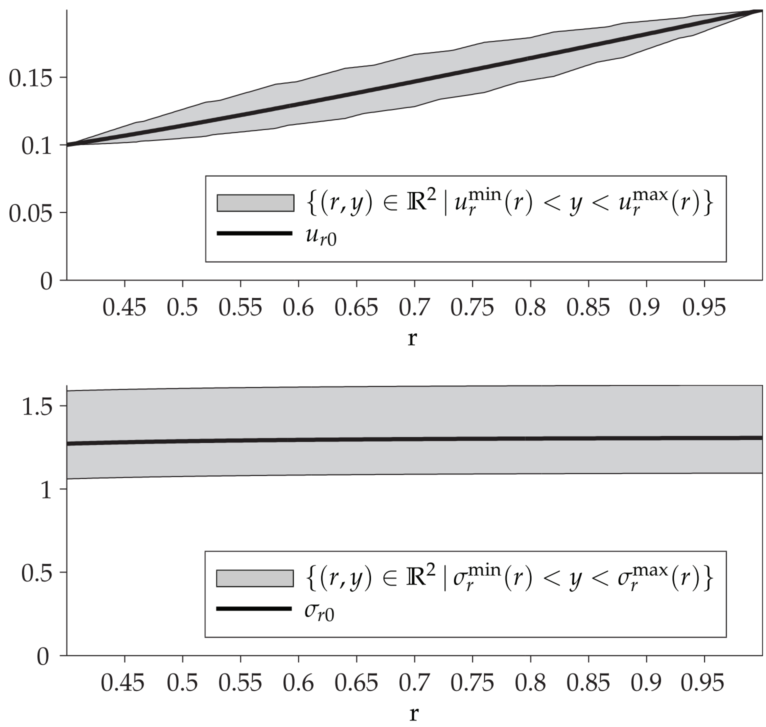

FIGURE 16 The set $\mathcal{S}(\mathcal{D}_h)$ and the respective mean solutions for the problem with pure Dirichlet conditions, $\nu_\circ = 0.45$, and $\delta = 0.01$.

energy,

$$e^J := \max_{u_{\text{pert}} \in \mathcal{S}(\mathcal{D}_h)} \frac{|J(u_{\text{pert}}) - J(u_\circ)|}{J(u_\circ)}.$$

These quantities are computed analytically on each interval. In all following experiments, we selected the interval to be $(a, b) := (0.4, 1.0)$, Young's modulus was set to one, Dirichlet conditions were $g_a = 0.01$ and $g_b = 0.02$, and the Neumann condition was $F_r = 0.2$. The perturbation set parameters were $N = 10$ and $M = 2$. In Figure 16 displacements and stresses of the set $\mathcal{S}(\mathcal{D}_h)$ are presented together with the non-perturbed mean solution for a particular problem.

In Figures 17 and 18, the development of various errors in the case of both boundary conditions is computed as a function of the relative perturbation magnitude $\frac{\delta}{\nu_\circ}$. From these plots one can depict the accuracy required for Poisson's ratio with respect to the desired accuracy and relevant quantity.

From Figures 17 and 18 we note that in every test example the accuracy required for computing point–wise values (L_∞-norm) is considerably higher than the accuracy required to compute the integral quantities. Moreover, in the case of the Neumann boundary conditions the quantities of interest are practically unaffected by the increased value of $\nu_\circ$, but in the case of the Dirichlet conditions they

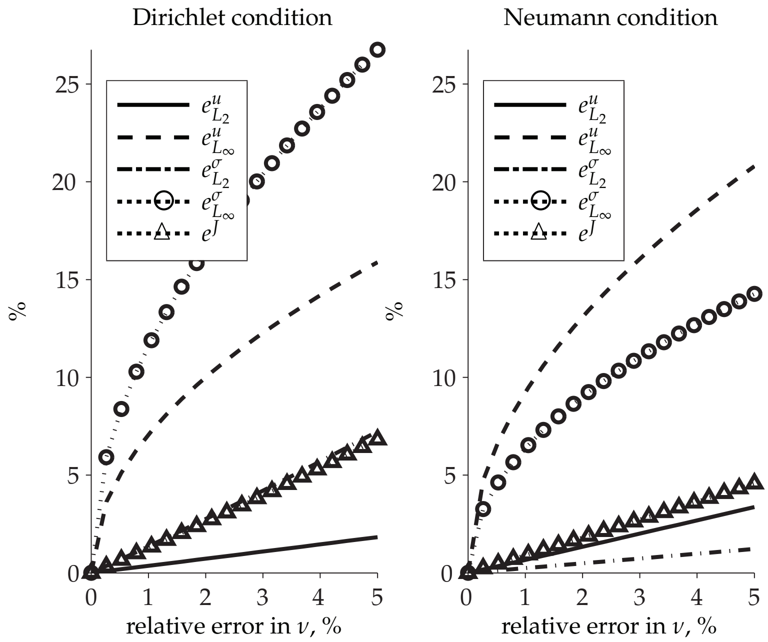

FIGURE 17 Relative perturbations of various quantities with respect to the relative perturbation in Poisson's ratio, when $\nu_\circ = 0.3$.

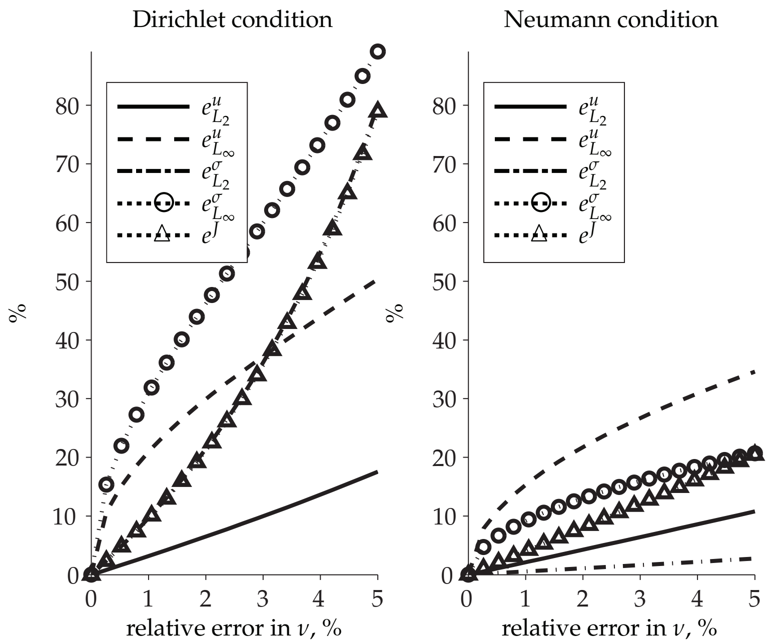

FIGURE 18 Relative perturbations of various quantities with respect to the relative perturbation in Poisson's ratio, when $\nu_\circ = 0.45$.

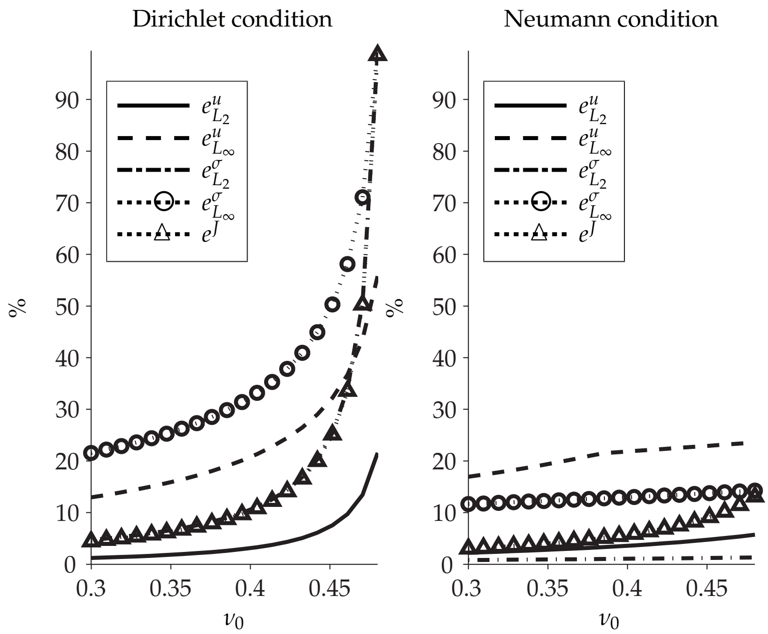

FIGURE 19 Relative perturbations of various quantities with respect to the mean Poisson's ratio $\nu_\circ$ when $\delta = 0.01$.

become very sensitive as ν increases. It is important to realize that this increased sensitivity of solutions to the variations of Poisson's ratio is highly relevant in engineering applications already before the neighbourhood of the incompressibility limit. This fact is highlighted in Figure 19, where we present the behaviour of relative perturbations with different $\nu_\circ$ and fixed δ for both boundary conditions.

5 CONCLUSIONS

An incomplete knowledge of the data in a model implies a limited accuracy of the results that the model generates. In order to perform a reliable quantitative analysis it is necessary to understand the effects generated by the incompletely known data. Moreover, these effects have to be controlled in a quantitative fashion with respect to energy error norms or other quantities of interest. In the Thesis, our main aim is to investigate the set formed by solutions associated with various possible data.

Theorem 4.1 provides two–sided bounds of the radius of the solution set for a certain class of elliptic problems. These bounds depend only on the known data of the problem and the magnitude of perturbations. We demonstrate that the derived upper bound is sharp. Theorem 4.1 shows that the perturbations around some given “mean” operator affect the solution (in the energy norm) almost proportionally. For linear elliptic problems this result is quite natural.

Another main result of the Thesis is Theorem 4.4. It shows that with a certain type of boundary conditions the sensitivity of the energy of the solution of the isotropic linear elasticity problem to Poisson’s ratio blows up. This behavior is demonstrated by a model problem that has an analytical solution. The asymptotic properties of the energy at the incompressibility limit indicate that a quantitative analysis in certain cases may face serious difficulties.

Our analysis suggests a method able to find concrete values of indeterminacy errors, which are important for quantitative (numerical) analysis of boundary value problems. The thesis is concerned with linear elliptic models. However, the method developed can, in principle, be extended to indeterminacy analysis of more complicated models if the functional deviation estimates are derived for them. One can expect that for more complicated problems the effects of the incomplete knowledge may be much more dramatic and imply serious consequences to the simulation practice.

REFERENCES

[1] M. Ainsworth and J. T. Oden, *A posteriori error estimation in finite element analysis*, Wiley and Sons, New york, 2000.

[2] I. Anjam, O. Mali, A. Muzalevsky, P. Neittaanmäki, and S. Repin, *A posteriori error estimates for a Maxwell type problem*, Russian J. Numer. Anal. Math. Modelling 24, no. 5, 2009.

[3] I. Babuška, *On randomized solutions of Laplace's equation*, Časopis Pěst. Mat. , Vol. 86, No. 3, pp. 269-276, 1961.

[4] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, John Wiley & Sons, 1997.

[5] A. Bernardini, *What are random and fuzzy sets and how to use them for uncertainty modelling in engineering systems?*, Whys and Hows in Uncertainty Modelling, Probability, Fuzziness and Anti-Optimization, edited I. Elishakoff, CISM Courses and Lectures, No. 388, Springer–Verlag, 1999.

[6] P. G. Ciarlet, *Finite element method for elliptic problems*, Studies in Mathematics and it's applications, Vol 4., North-Holland, Amsterdam, 1978.

[7] A. P. Dempster, *Upper and lower probabilities induced by a multi–valued mapping*, Ann. Math. Statist. 38, pp. 325–339, 1967.

[8] G. Duvaut and J.-L. Lions, *Les inequations en mecanique et en physique*, Dunod, Paris, 1972.

[9] I. Ekeland and R. Teman, *Convex analysis and variational problems*, North–Holland, Amsterdam 1976.

[10] I. Elishakoff, *Probabilistic methods in the Theory of Structures*, J. Wiley, New York, 1983.

[11] M. Frolov, *Reliable control over approximation errors by functional type a posteriori error estimates*, PhD thesis, Jyväskylä Studies in Computing 44, University of Jyväskylä, 2004.

[12] M. Frolov, P. Neittaanmäki, and S. Repin, *On computational properties of a posteriori error estimates based upon the method of duality error majorants*, Numerical mathematics and advanced applications, Springer–Verlag, Berlin, 2000.

[13] E. Gorshkova, *A posteriori error estimates and adaptive methods for incompressible viscous flow problems*, PhD thesis, Jyväskylä Studies in Computing 86, University of Jyväskylä, 2007.

[14] E. Gorshkova, A. Mahalov, P. Neittaanmäki, and S Repin, *Functional a posteriori estimates for viscous flow problems with rotation*, J. Math. Sci. New York, 142, pp. 927–935, 2007.

[15] A. Hannukainen, *Functional a posteriori error estimates for the Maxwell's equations*, ENUMATH 2007, September 10–14, book of abstracts, p. 114, 2007.

[16] E. J. Haug, K. K. Choi, and V. Komkov, *Design Sensitivity Analysis of Structural Systems*, Academic Press, Orlando, 1986.

[17] J. Haslinger and P. Neittaanmäki, *Finite element approximation for optimal shape, material and topology design*, Second edition, John Wiley & Sons, Ltd., Chichester, 1996.

[18] J. Haslinger and R. Mäkinen, *Introduction to shape optimization. Theory, approximation, and computation*, Advances in Design and Control, 7. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2003.

[19] I. Hlaváček, J. Chleboun, and I. Babuška, *Uncertain input data problems and the worst scenario method*, Elsevier, Amsterdam, 2004.

[20] J. P. C. Klejnen and J. C. Helton, *Statistical analyses of scatterplots to identify important factors in large-scale simulations, 1: Review and comparison of techniques*, Reliab. Engng. Syst. Safety, 65, pp. 147–185, 1999.

[21] J. P. C. Klejnen and J. C. Helton, *Statistical analyses of scatterplots to identify important factors in large-scale simulations, 2: robustness of techniques*, Reliab. Engng. Syst. Safety, 65, pp. 187–197, 1999.

[22] O. A. Ladyzenskaja and N. N. Uraltseva, *Linear and Quasilinear Elliptic equations*, Academic Press, New York, 1968.

[23] V. G. Litvinov, *Optimization in Elliptic Boundary Value Problems with Applications to Mechanics of Deformable Bodies and Fluid Mechanics*, Birkhaäuser, Berlin, 2000.

[24] O. Mali, *A Posteriori Error Estimates for the Kirchhoff-Love Arch Model*, Proceedings of the 10th Finnish Mechanics Days, ed. R. Mäkinen, P. Neittaanmäki, T. Tuovinen, and K. Valpe, Reports of the Department of Mathematical Information Technology Series A. Collections, No. A 1/2009, pp. 315-323, 2009.

[25] O. Mali and S. Repin, *Error control for problems with uncertain data by functional a posteriori estimates*, In "Adaptive Modeling and Simulation 2007", ed. K. Runesson and P. Diez, CIMNE, Barcelona, pp. 158-161, 2007.

[26] O. Mali and S. Repin, *Estimates of the indeterminacy set for elliptic boundary–value problems with uncertain data, J. Math. Sci.* 150, pp. 1869-1874, 2008.

[27] O. Mali and S. Repin, *Estimates of accuracy limit for elliptic boundary value problems with uncertain data*, Advances in Mathematical Sciences and Applications 19, pp. 525-537, 2009.

[28] O. Mali and S.Repin, *Two-sided estimates of the solution set for the reaction-diffusion problem with uncertain data*, Applied and numerical partial differential equations, 183–198, Comput. Methods Appl. Sci., 15, Springer, New York, 2010.

[29] O. Mali and S.Repin, *Blowup of the energy increment caused by uncertainty of the Poisson' ratio in elasticity problems*, Russian J. Numer. Anal. Math. Modelling 26, no. 4, pp. 413-425, 2011.

[30] S. G. Mikhlin, *Error Analysis in Numerical Processes*, Wiley and Sons, Chester – New York, 1991.

[31] P. Neittaanmäki, J. Sprekels, and D. Tiba, *Optimization of elliptic systems, Theory and applications*, Springer Monographs in Mathematics. Springer, New York, 2006.

[32] P. Neittaanmäki and S. Repin: *Reliable Methods for Computer Simulation, Error Control and A Posteriori Estimates*, Elsevier, 2004.

[33] W. Prager and J. L. Synge, *Approximation in elasticity based on the concept of function space*, Quart. Appl. Math., Vol. 5, pp. 241–269, 1947.

[34] R. Rannacher, *The dual-weighted-residual method for error control and mesh adaptation schemes in finite element methods*, The mathematics of finite elements and applications X, MAEFLAP 1999. Proceedings of the 10th conference, Brunel Univ., Uxbridge, Middlesex, GB, June 22–25, 1999, edited by J. Whiteman, Elsevier, Amsterdam, pp. 97–116, 2000.

[35] S. Repin, *A posteriori error estimation for nonlinear variational problems with power growth functionals based on duality theory*, Zapiski Nauchn, Semin, V. A. Steklov Mathematical Institute in St. Petersburg (POMI) 249, pp. 244–255, 1997.

[36] S. Repin, *A posteriori error estimation for variational problems with uniformly convex functionals*, Math. Comput. 69, pp. 481-500, 2000.

[37] S. Repin, *Two-sided estimates of deviation from exact solutions of uniformly elliptic equations,* Proc. St. Petersburg Math. Society, IX(2001), pp. 143–171, translation in Amer. Math. Soc. Transl. Ser. 2, 209, Amer. Math. Soc., Providence, RI, 2003.

[38] S. Repin, *A posteriori error estimates taking into account indeterminacy of the problem data,* Russian J. Numer. Anal. Math. Modelling, 18, pp. 507–519, 2003.

[39] S. Repin, *A Posteriori Estimates for Partial Differential Equations*, Walter de Gruyter, 2008.

[40] S. Repin and S. Sauter, *Functional a posteriori estimates for the reaction diffusion problem*, C. R. Acad. Sci. Paris, Sér. I, Math 343, pp. 349-354, 2006.

[41] S. Repin and S. Sauter, *Estimates of the modeling error for the Kirchhoff-Love plate model*, C. R. Math. Acad. Sci. Paris 348, no. 17-18, pp. 1039–1043, 2010.

[42] S. Repin, S Sauter, and A. Smolianski, *A posteriori estimation of dimension reduction errors for elliptic problems on thin domains*, SIAM J. Numer. Anal. 42, no. 4, pp. 1435–1451, 2004.

[43] S. Repin, S. Sauter, and A. Smolianski, *A posteriori error estimation for the Dirichlet problem with account of the error in the approximation of boundary conditions*, Computing 70, no. 3, pp. 205–233, 2003.

[44] S. Repin and R. Stenberg, *A posteriori error estimates for the generalized Stokes problem*, Journal of Mathematical Sciences, Vol. 142, No. 1, 2007.

[45] T. Roubíček, *Relaxation in Optimization Theory and Variational Calculus*, De Gruyter Series in Nonlinear Analysis and Applications, Walter de Gruyter, Berlin, 1997.

[46] G. I. Schuëller, *A state-of-the-art report on computational stochastic mechanics*, Prob. Engrg. Mech, Vol 12, No. 4, pp. 197-321, 1997.

[47] G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, 1976.

[48] S. L. Sobolev, *Some Applications of Functionals Analysis in Mathematical Physics*, Izdt, Leningrad. Gos. Univ., Leningrad, 1955 (in Russian, translated in Translation of Mathematical Monographs, Volume 90, AMS, Providence, RI 1991).

[49] P. Solin, *Partial Differential Equations and the Finite Element Method*, John Wiley & Sons, 2006.

[50] R. Stenberg, *Error analysis of some finite element methods for the Stokes problem*, Math. Comput. 54, pp. 495–508, 1990.

[51] D. Tiba and R. Vodak, *A General Asymptotic Model for Lipschitzian Curved Rods*, Advances in Mathematical Sciences and Applications, Vol. 15, No. 1, pp. 137-198, 2005.

[52] S. Timoshenko and J. N. Goodier, *Elasticity theory*, McGraw-Hill, New York, 1951.

[53] S. Timoshenko and D. H. Young, *Theory of Structures*, New York, McGraw-Hill Book Company, Inc., 1945.

[54] J. Valdman, *Minimization of Functional Majorant in A Posteriori Error Analysis Based on H(div) Multigrid–Preconditioned CG Method*, Advances in Numerical Analysis, Article ID 164519, Volume 2009.

[55] R. Verfurth, *A review of a posteriori error estimation and adaptive mesh refinement techniques*, Wiley and Sons, Teubner–Verlag, New York, 1996.

[56] L. A. Zadeh, *Fuzzy sets*, Information and control, 8. pp. 338-353, 1965 .

[57] L. A. Zadeh, *Fuzzy sets as basis for theory of possibility*, Fuzzy sets and systems, 1. pp. 3-28, 1978.

[58] O. C. Zienkiewits and J. Z. Zhu, *A simple error estimator and adaptive procedure for practical engineering analysis*, Internat. J. Numer. Meth. Engrg, 24, pp. 337–357, 1987.

[59] E. Zio, *Reliability engineering: Old problems and new challenges*, Reliability Engineering and System Safety 94, pp. 125-141, 2009.

YHTEENVETO (FINNISH SUMMARY)

Insinöörisovelluksissa ja luonnontieteissä laajemminkin matemaattisten mallien avulla pyritään ennustamaan ilmöitä ja tapahtumia. On tärkeää ymmärtää, että mikäli lähtötiedot tunnetaan vain rajallisella tarkkuudella, ei niistä voida johtaa mielivaltaisen tarkkoja ennusteita. On jopa mahdollista, että mallinnettava ilmiö on herkkä jonkun parametrin suhteen, eli pienetkin muutokset lähtöarvoissa voivat aiheuttaa merkittäviä poikkeamia tuloksissa.

Tässä väitöskirjassa tutkitaan puutteellisesti tunnettujen parametrien vaikutusta yhtälön ratkaisuun. Usein matemaattiset mallit sisältävät parametreja, joita ei tunneta tarkasti. Tilanne on tyypillinen esimerkiksi kontinuumimekaniikassa, jossa materiaalien konstitutiiviset relaatiot perustuvat oletuksiin ja kokeellisiin tuloksiin. Tässä työssä tarkastellut mallit ovat erään tyypin elliptisiä osittaisdifferentiaaliyhtälöitä. Monet fysikaalisten ilmiöiden tasapainotilaa kuvaavat matemaattiset mallit ovat tätä tyyppiä, esimerkiksi lämmönjohtavuus, diffuusio, magnetostatiikka ja erityisesti lineaarielastiikka. Vaikka väitöskirjan otsikko, "Puutteellisesti tunnetun materiaalikäyttäymisen aiheuttaman virheen analyysi elastisen aineen matemaattisissa malleissa", alleviivaa tutkimuksen yhteyttä juuri lineaarielastiikan malleihin, ovat tulokset suoraan hyödynnettävissä myös muiden mainittujen fysiikan ilmiöiden mallinnuksessa.

Suoritetun analyysin keskeisimpinä työkaluina käytetään funktionaalisia a posteriori estimaatteja. Ne ovat funktionaaleja, joilla voidaan arvoida minkä tahansa käypän funktion etäisyyttä mallin tarkasta ratkaisusta. Lisäksi ne riippuvat eksplisiittisesti ongelman parametreista. Tässä työssä johdetaan kyseiset estimaatit myös kaarevia palkkeja käsittelevälle Kirchhoff–Love palkkimallille.

Tehdyn tutkimuksen päätavoite on tarjota asiantuntijoille työkaluja ja metodeja, joilla voidaan saada vastaus kahteen mallintamiseen kannalta keskeiseen kysymykseen:

(a) Kun lähtöarvot tunnetaan tietyllä tarkkuudella, kuinka suuri tarkkuus simulaatiotuloksissa on mahdollista saavuttaa?

(b) Jotta tulokset tunnettaisiin halutulla tarkkuudella, kuinka suuri epätarkkuus lähtöarvoissa voidaan sallia?

Tässä työssä arvioidaan ratkaisujoukon, eli mahdollisten lähtöarvojen tuottamien ratkaisujen joukon, kokoa (sädettä) ala- ja yläpuolelta. Johdetut arviot ratkaisujoukon koosta riippuvat ainoastaan tutkittavan ongelman ominaisuuksista ja parametreissa esiintyvän epävarmuuden suuruudesta. Yksinkertainen esimerkkitehtävä osoittaa, että johdettu yläraja on tarkka. Toisin sanoen on olemassa tehtäviä, joiden ratkaisujoukko kasvaa täsmälleen funktionaalisten estimaattien avulla johdetun ylärajan osoittamalla tavalla, kun lähtötietojen epätarkkuutta kasvatetaan.

Toinen väitöskirjan keskeinen tulos on lineaarielastiikan ongelman ratkaisun kokonaisenergian tarkastelu, kun Poissonin luku oletetaan puutteellisesti tunnetuksi. Tehtävän reunaehtotyypit (käypien ratkaisujen joukko) voidaan jakaa kahteen luokkaan, joiden kvalitatiivinen käytös poikkeaa toisistaan huomattavasti kun materiaali on lähes kokoonpuristumatonta (Poissonin luku on lähes puoli). Tulos on luonteeltaan kvalitatiivinen eikä suoraan kerro ratkaisujoukon koosta, vaan sen kasvunopeudesta Poissonin luvun suhteen. On syytä huomata, että jo huomattavasti ennen asymptoottista kokoonpuristumattomuusrajaa, puutteellisesti tunnetun Poissonin luvun aiheuttama epävarmuus ratkaisussa voi tehdä kaiken kvantitatiivisen analyysin liian epätarkaksi insinöörisovellusten kannalta. Epävarmuuden vaikutuksia havainnollistetaan analyyttisesti ratkeavan esimerkkitehtävän avulla.

Väitöskirjassa käsitellään vain lineaarisia elliptisiä tehtäviä. Teoriassa samankaltainen analyysi voidaan suorittaa myös monimutkaisemmille tehtäville, mikäli vastaavat funktionaaliset a posteriori estimaatit tunnetaan. On odotettavissa, että puutteellisesti tunnettujen lähtötietojen vaikutukset ovat vielä huomattavampia monimutkaisemmille tehtäville ja niiden huomioonottamisella on tulee olemaan merkittävä vaikutus mallinnuskäytäntöihin.

J Y V Ä S K Y L Ä S T U D I E S I N C O M P U T I N G

1 ROPPONEN, JANNE, Software risk management - foundations, principles and empirical findings. 273 p. Yhteenveto 1 p. 1999.
2 KUZMIN, DMITRI, Numerical simulation of reactive bubbly flows. 110 p. Yhteenveto 1 p. 1999.
3 KARSTEN, HELENA, Weaving tapestry: collaborative information technology and organisational change. 266 p. Yhteenveto 3 p. 2000.
4 KOSKINEN, JUSSI, Automated transient hypertext support for software maintenance. 98 p. (250 p.) Yhteenveto 1 p. 2000.
5 RISTANIEMI, TAPANI, Synchronization and blind signal processing in CDMA systems. - Synkronointi ja sokea signaalinkäsittely CDMA järjestelmässä. 112 p. Yhteenveto 1 p. 2000.
6 LAITINEN, MIKA, Mathematical modelling of conductive-radiative heat transfer. 20 p. (108 p.) Yhteenveto 1 p. 2000.
7 KOSKINEN, MINNA, Process metamodelling. Conceptual foundations and application. 213 p. Yhteenveto 1 p. 2000.
8 SMOLIANSKI, ANTON, Numerical modeling of two-fluid interfacial flows. 109 p. Yhteenveto 1 p. 2001.
9 NAHAR, NAZMUN, Information technology supported technology transfer process. A multi-site case study of high-tech enterprises. 377 p. Yhteenveto 3 p. 2001.
10 FOMIN, VLADISLAV V., The process of standard making. The case of cellular mobile telephony. - Standardin kehittämisen prosessi. Tapaustutkimus solukkoverkkoon perustuvasta matkapuhelintekniikasta. 107 p. (208 p.) Yhteenveto 1 p. 2001.
11 PÄIVÄRINTA, TERO, A genre-based approach to developing electronic document management in the organization. 190 p. Yhteenveto 1 p. 2001.
12 HÄKKINEN, ERKKI, Design, implementation and evaluation of neural data analysis environment. 229 p. Yhteenveto 1 p. 2001.
13 HIRVONEN, KULLERVO, Towards better employment using adaptive control of labour costs of an enterprise. 118 p. Yhteenveto 4 p. 2001.
14 MAJAVA, KIRSI, Optimization-based techniques for image restoration. 27 p. (142 p.) Yhteenveto 1 p. 2001.
15 SAARINEN, KARI, Near infra-red measurement based control system for thermo-mechanical refiners. 84 p. (186 p.) Yhteenveto 1 p. 2001.
16 FORSELL, MARKO, Improving component reuse in software development. 169 p. Yhteenveto 1 p. 2002.
17 VIRTANEN, PAULI, Neuro-fuzzy expert systems in financial and control engineering. 245 p. Yhteenveto 1 p. 2002.
18 KOVALAINEN, MIKKO, Computer mediated organizational memory for process control. Moving CSCW research from an idea to a product. 57 p. (146 p.) Yhteenveto 4 p. 2002.
19 HÄMÄLÄINEN, TIMO, Broadband network quality of service and pricing. 140 p. Yhteenveto 1 p. 2002.
20 MARTIKAINEN, JANNE, Efficient solvers for discretized elliptic vector-valued problems. 25 p. (109 p.) Yhteenveto 1 p. 2002.
21 MURSU, ANJA, Information systems development in developing countries. Risk management and sustainability analysis in Nigerian software companies. 296 p. Yhteenveto 3 p. 2002.
22 SELEZNYOV, ALEXANDR, An anomaly intrusion detection system based on intelligent user recognition. 186 p. Yhteenveto 3 p. 2002.
23 LENSU, ANSSI, Computationally intelligent methods for qualitative data analysis. 57 p. (180 p.) Yhteenveto 1 p. 2002.
24 RYABOV, VLADIMIR, Handling imperfect temporal relations. 75 p. (145 p.) Yhteenveto 2 p. 2002.
25 TSYMBAL, ALEXEY, Dynamic integration of data mining methods in knowledge discovery systems. 69 p. (170 p.) Yhteenveto 2 p. 2002.
26 AKIMOV, VLADIMIR, Domain decomposition methods for the problems with boundary layers. 30 p. (84 p.). Yhteenveto 1 p. 2002.
27 SEYUKOVA-RIVKIND, LUDMILA, Mathematical and numerical analysis of boundary value problems for fluid flow. 30 p. (126 p.) Yhteenveto 1 p. 2002.
28 HÄMÄLÄINEN, SEPPO, WCDMA Radio network performance. 235 p. Yhteenveto 2 p. 2003.
29 PEKKOLA, SAMULI, Multiple media in group work. Emphasising individual users in distributed and real-time CSCW systems. 210 p. Yhteenveto 2 p. 2003.
30 MARKKULA, JOUNI, Geographic personal data, its privacy protection and prospects in a location-based service environment. 109 p. Yhteenveto 2 p. 2003.
31 HONKARANTA, ANNE, From genres to content analysis. Experiences from four case organizations. 90 p. (154 p.) Yhteenveto 1 p. 2003.
32 RAITAMÄKI, JOUNI, An approach to linguistic pattern recognition using fuzzy systems. 169 p. Yhteenveto 1 p. 2003.
33 SAALASTI, SAMI, Neural networks for heart rate time series analysis. 192 p. Yhteenveto 5 p. 2003.
34 NIEMELÄ, MARKETTA, Visual search in graphical interfaces: a user psychological approach. 61 p. (148 p.) Yhteenveto 1 p. 2003.
35 YOU, YU, Situation Awareness on the world wide web. 171 p. Yhteenveto 2 p. 2004.
36 TAATILA, VESA, The concept of organizational competence – A foundational analysis. - Perusteanalyysi organisaation kompetenssin käsitteestä. 111 p. Yhteenveto 2 p. 2004.

106 VASILYEVA, EKATERINA, Tailoring of feedback in web-based learning systems: Certitude-based assessment with online multiple choice questions. 124 p. (184 p.) Yhteenveto 2 p. 2009.

107 KUDRYASHOVA, ELENAV., Cycles in continuous and discrete dynamical systems. Computations, computer assisted proofs, and computer experiments. 79 p. (152 p.) Yhteenveto 1 p. 2009.

108 BLACKLEDGE, JONATHAN, Electromagnetic scattering and inverse scattering solutions for the analysis and processing of digital signals and images. 297 p. Yhteenveto 1 p. 2009.

109 IVANNIKOV, ANDRIY, Extraction of event-related potentials from electroencephalography data. - Herätepotentiaalien laskennallinen eristäminen EEG-havaintoaineistosta. 108 p. (150 p.) Yhteenveto 1 p. 2009.

110 KALYAKIN, IGOR, Extraction of mismatch negativity from electroencephalography data. - Poikkeavuusnegatiivisuuden erottaminen EEG-signaalista. 47 p. (156 p.) Yhteenveto 1 p. 2010.

111 HEIKKILÄ, MARIKKA, Coordination of complex operations over organisational boundaries. 265 p. Yhteenveto 3 p. 2010.

112 FEKETE, GÁBOR, Network interface management in mobile and multihomed nodes. 94 p. (175 p.) Yhteenveto 1 p. 2010.

113 KUJALA, TUOMO, Capacity, workload and mental contents - Exploring the foundations of driver distraction. 146 p. (253 p.) Yhteenveto 2 p. 2010.

114 LUGANO, GIUSEPPE, Digital community design - Exploring the role of mobile social software in the process of digital convergence. 253 p. (316 p.) Yhteenveto 4 p. 2010.

115 KAMPYLIS, PANAGIOTIS, Fostering creative thinking. The role of primary teachers. - Luovaa ajattelua kehittämässä. Alakoulun opettajien rooli. 136 p. (268 p.) Yhteenveto 2 p. 2010.

116 TOIVANEN, JUKKA, Shape optimization utilizing consistent sensitivities. - Muodon optimointi käyttäen konsistentteja herkkyyksiä. 55 p. (130p.) Yhteenveto 1 p. 2010.

117 MATTILA, KEIJO, Implementation techniques for the lattice Boltzmann method. - Virtausdynamiikan tietokonesimulaatioita Hila-Boltzmann -menetelmällä: implementointi ja reunaehdot. 177 p. (233 p.) Yhteenveto 1 p. 2010.

118 CONG, FENGYU, Evaluation and extraction of mismatch negativity through exploiting temporal, spectral, time-frequency, and spatial features. - Poikkeavuusnegatiivisuuden (MMN) erottaminen aivosähkönauhoituksista käyttäen ajallisia, spektraalisia, aikataajuus - ja tilapiirteitä. 57 p. (173 p.) Yhteenveto 1 p. 2010.

119 LIU, SHENGHUA, Interacting with intelligent agents. Key issues in agent-based decision support system design. 90 p. (143 p.) Yhteenveto 2 p. 2010.

120 AIRAKSINEN, TUOMAS, Numerical methods for acoustics and noise control. - Laskennallisia menetelmiä akustisiin ongelmiin ja melunvaimennukseen. 58 p. (133 p.) Yhteenveto 2 p. 2010.

121 WEBER, MATTHIEU, Parallel global optimization Structuring populations in differential evolution. - Rinnakkainen globaalioptimointi. Populaation rakenteen määrittäminen differentiaalievoluutiossa. 70 p. (185 p.) Yhteenveto 2 p. 2010.

122 VÄÄRÄMÄKI, TAPIO, Next generation networks, mobility management and appliances in intelligent transport systems. - Seuraavan sukupolven tietoverkot, liikkuvuuden hallinta ja sovellutukset älykkäässä liikenteessä. 50 p. (111 p.) Yhteenveto 1 p. 2010.

123 VIUKARI, LEENA, Tieto- ja viestintätekniikkavälitteisen palvelun kehittämisen kolme diskurssia. - Three discourses for an ICT-service development . 304 p. Summary 5 p. 2010.

124 PUURTINEN, TUOMAS, Numerical simulation of low temperature thermal conductance of corrugated nanofibers. - Poimutettujen nanokuitujen lämmönjohtavuuden numeerinen simulointi matalissa lämpötiloissa . 114 p. Yhteenveto 1 p. 2010.

125 HILTUNEN, LEENA, Enhancing web course design using action research . - Verkko-opetuksen suunnittelun kehittäminen toimintatutkimuksen keinoin . 192 p. Yhteenveto 2 p. 2010.

126 AHO, KARI, Enhancing system level performance of third generation cellular networks through VoIP and MBMS services. 121 p. (221 p.). Yhteenveto 2 p. 2010.

127 HÄKKINEN, MARKKU, Why alarms fail. A cognitive explanatory model. 102 p. (210 p.). Yhteenveto 1 p. 2010.

128 PENNANEN, ANSSI, A graph-based multigrid with applications. - Graafipohjainen monihilamenetelmä sovelluksineen. 52 p. (128 p.). Yhteenveto 2 p. 2010.

129 AHLGREN, RIIKKA, Software patterns, organizational learning and software process improvement. 70 p. (137 p.). Yhteenveto 1 p. 2011.

130 NIKITIN, SERGIY, Dynamic aspects of industrial middleware architectures 52 p. (114 p.). Yhteenveto 1 p. 2011.

131 SINDHYA, KARTHIK, Hybrid Evolutionary Multi-Objective Optimization with Enhanced Convergence and Diversity. 64 p. (160 p.). Yhteenveto 1 p. 2011.