PDE-constrained optimization: Preconditioners and diffuse domain methods

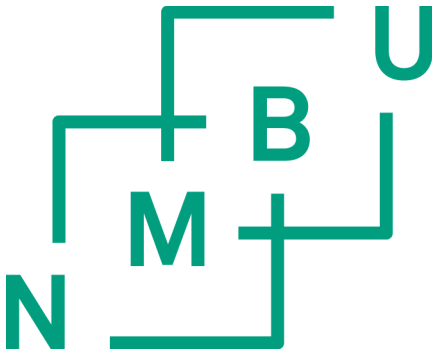
PDE-betinga optimering:
Prekondisjonerarar og metodar for diffuse domene

Philosophiae Doctor (PhD) Thesis

Ole Løseth Elvetun

Department of Mathematical Sciences and Technology
Faculty of Environmental Science and Technology
Norwegian University of Life Sciences

Ås (2015)



Thesis number 2015:84
ISSN 1894-6402
ISBN 978-82-575-1319-1

Abstract

This thesis is mainly concerned with the efficient numerical solution of optimization problems subject to linear PDE-constraints, with particular focus on robust preconditioners and diffuse domain methods. Associated with such constrained optimization problems are the famous first-order Karush-Kuhn-Tucker (KKT) conditions. For certain minimization problems, the functions satisfying the KKT conditions are also optimal solutions of the original optimization problem, implying that we can solve the KKT system to obtain the optimum; the so-called "all-at-once" approach. We propose and analyze preconditioners for the different KKT systems we derive in this thesis.

In papers I and II we study PDE-constrained optimization problems with inequality constraints and problems subject to total variation regularization, respectively. These are both non-linear problems, so we apply iterative methods; the Primal Dual Active Set algorithm and the split Bregman method, resulting in iterative schemes where we must solve a sequence of linear KKT systems. Using Riesz maps to form preconditioners, we get iteration numbers independent of the mesh parameter h, and we are able to prove a maximum growth in MINRES iteration numbers of order $O([\log(\alpha^{-1})]^2)$ as the regularization parameter $\alpha \to 0$. Furthermore, we present numerical simulations with the improved rate of order $O(\log(\alpha^{-1}))$.

To derive a solver which is completely robust with respect to both the mesh parameter h and the regularization parameter α is, from a functional analysis perspective, a matter of finding weighted Sobolev spaces in which all the stability estimates are independent of h and α. If such topologies are obtained, the Riesz maps associated with the underlying normed spaces will form a natural preconditioner for the KKT system, resulting in solvers with h- and α-independent iteration numbers.

The third paper concerns the derivation of such a robust preconditioner for a specific PDE-constrained optimization problem. More specifically, we

study an elliptic control problem with boundary observations only and locally defined control functions. A careful analysis reveals that there exists an isomorphism between the control space and the space of Lagrange multipliers, leading to stability estimates of the associated KKT system independent of the mesh parameter h and regularization parameter α. Consequently, we obtain a completely h- and α-robust Krylov subspace solver. The problem studied in Paper III was motivated by the inverse problem of electrocardiography (ECG).

Finally, in papers IV and V, we are concerned with the computational representation of the involved domains. In applications, the domains are often complex or not exactly known. We apply the diffuse domain method, an embedding technique, to solve PDE-constrained optimization problems posed on such domains. A full theoretical investigation is undertaken, and strict convergence rates, with respect to the diffuse domain parameter ϵ, is obtained. We must also handle topologies depending on the parameter ϵ, which increases the complexity of deriving robust KKT solvers. A completely ϵ-robust iterative solver is, nevertheless, achieved from a careful construction of topologies.

All the theoretical investigations, presented in this thesis, are supported by numerical simulations, and we obtain very good agreement between the theoretical and numerical results.

Samandrag

Denne avhandlinga ser i hovudsak på effektive numeriske løysingar av PDE-betinga optimeringsproblem, med eit særskilt fokus på robuste prekondisjonerar og "diffuse domain"-metodar. Assosiert med slike optimeringsproblem er dei velkjende Karush-Kuhn-Tucker (KKT)-føresetnadane. For mange betinga optimeringsproblem, vil funksjonar som tilfredstillar KKT-vilkåra samstundes vere ei optimal løysing på det opprinnelege optimeringsproblemet. Dette impliserar at vi kan løyse KKT-likningane for å finne optimum. Vi konstruerar og analyserar prekondisjonerar for dei forskjellige KKT-systema vi utleiar i denne avhandlinga.

I artikkel I studerar vi kontrollproblem med ulikskapsvilkår på kontrollfunksjonen, medan vi i artikkel II analyserar optimeringsproblem underlagt totalvariasjonsregularisering. Begge desse problema er ikkje-lineære, som gjer at vi må nytte iterative metodar for å løyse problema. Ved bruk av hhv. "the Primal Dual Active Set"- og "split Bregman"- algoritmen, får vi iterative skjema kor vi må løyse ein sekvens av lineære KKT-system. Brukar vi Riesz-operatorar til å danne prekondisjonerar, får vi iterasjonstal som er uavhengige av meshparameteren h og vi beviser ein maksimal vekst i iterasjonstal av orden $O([\log(\alpha^{-1})]^2)$ når regulariseringsparameteren $\alpha \to 0$. I tillegg syner vi numeriske simuleringar med den forbetra raten $O(\log(\alpha^{-1}))$.

Å finne ein løysar som er heilt robust med omsyn på både meshparameteren h og regulariseringsparameteren α er, i frå eit funksjonalanalyseperspektiv, eit spørsmål om å finne vekta Sobolevrom der alle stabilitetsestimata er uavhengige av h og α. Gitt slike topologiar, vil Riesz-operatorane som er assosiert med dei underliggande normerte romma danne ein naturleg prekondisjonerar for KKT-systemet.

Den tredje artikkelen omhandlar utleiinga av ein slik robust prekondisjonerar for eit spesifikt PDE-betinga optimeringsproblem. Vi studerar eit elliptisk kontrollproblem med kun randobservasjonar og ein lokalt definert

kontrollfunksjon. Ein rigorøs analyse synar at det eksisterar ein isomorfi mellom kontrollrommet og rommet av Lagrangemultiplikatorar, som fører til stabilitetsestimat som er uavhengige av h og α. Følgeleg får vi ein fullstendig h- og α-robust Krylovromløysar.

Til slutt, i artikkel IV og V, er vi oppteken av numerisk representasjonen av dei involverte domena. For reelle problem er domena ofte komplekse eller ikkje nøyaktig kjende. Vi nyttar "diffuse domain"-metoden, ein embeddingsteknikk, for å løyse PDE-betinga optimeringsproblem gitt på slike domene. Ein full teoretisk analyse er gjennomført, og konvergensrater med omsyn på "diffuse domain"-parameteren ϵ er oppnådd. I tillegg handsamar vi her topologiar som er avhengige av parameteren ϵ, som gjer det meir utfordrande å oppnå robust prekondisjonering av KKT-systema. Ein ϵ-robust Krylovromløysar er likevel utleia frå ein nøye konstruksjon av dei involverte topologiane.

Alle dei teoretiske undersøkingane i avhandlinga er støtta av numeriske simuleringar, og vi oppnår veldig godt samsvar mellom dei teoreriske og numeriske resultata.

"For the things of this world cannot be made known without a knowledge of mathematics"
- Roger Bacon

Acknowledgements

I would first of all like to thank my main supervisor Professor Bjørn Fredrik Nielsen. From the very beginning of my PhD project, he has offered invaluable support and encouraged me to work focused and oriented towards the goal. Many a time have we discussed and solved both mathematical issues and other world problems.

The other members of the mathematics group at the department should also to be thanked. Their doors have been open for me and my questions throughout my years at the university.

As a part of my doctoral work I spent a year in Professor Martin Burger's Imaging Workgroup at the University of Münster. I would like to express my gratitude for the warm welcome I received from the entire group. Although Professor Burger is not among my official supervisors, he has provided much guidance and introduced me to interesting new branches of applied mathematics. He also connected me with Dr. Matthias Schlottbom, resulting in the three of us writing two papers together. It has been truly inspiring working with these two gifted mathematicians. I look back upon my stay in Münster fondly, only regretting having left Germany two weeks before the final of the 2014 World Cup.

I am also grateful for the support from my co-supervisor, Associate Professor Kent-Andre Mardal. Particularly for introducing me to the FEniCS software package for block operators. Without his work developing this library, I would probably still be writing code.

Finally, my family deserves my deepest gratitude. My parents for always having supported me and encouraged me to work hard to achieve my goals and my wife Anette Elvetun for the love and laughter we share. Thank you for always believing in me.

Ole Løseth Elvetun Ås, July 2015

Contents

1

Introduction

1.1 Background

1.1.1 Motivation

Mathematics is used to model phenomena in a broad range of scientific and industrial disciplines. Many of these models are formulated in terms of partial differential equations (PDEs). These equations can be used to simulate electrical potentials, heat conduction, groundwater flow, water waves, electromagnetic waves, etc.

A large class of PDEs is deterministic, i.e. there is no randomness in the model. This means that a given initial state will always produce the same output. In more philosophical terms, we might say that the equations model the effect (output) of a given cause (initial state/input data). In practical applications, we often have no information about the initial state, but we can only observe (parts of) the effect. This can be formulated as a PDE-constrained optimization problem: We search for the initial state, or source, which produces the best approximation of the measured output.

PDE-constrained optimization is an active research field, and there is a vast number of challenges to investigate. We will make no attempt to address all of these issues, but rather focus on those we study in this thesis. To motivate the particular choice of topics, we present the application which inspired our selection of problems: The inverse problem of electrocardiography.

The aim of this inverse problem is to identify an ischemic[1] region in the heart by combining ECG recordings with the bidomain model. We will

[1] Ischemia is a state of reduced blood supply to the heart, usually due to coronary artery disease. It is a reversible condition, but also a precursor to a heart infarct.

not go into details on how to derive the model, but rather refer to [9, 12]. Instead, we will present the involved PDE and motivate how we can use this model to locate an ischemic area.

The PDE reads: Find $u \in H^1(\Omega_B)$ such that

$$\int_{\Omega_B} \nabla\psi \cdot M\nabla u \, dx = -\int_{\Omega_H} \nabla\psi \cdot M_i \nabla f \, dx, \ \forall\psi \in H^1(\Omega_B), \tag{1.1}$$

where

1. $u \in H^1(\Omega_B)$ is the extracellular potential, i.e. the potential outside the heart cells,
2. $f \in H^1(\Omega_H)$ is the transmembrane potential, i.e. the potential difference over the cell membrane of the heart cells,
3. Ω_B is the domain of the body (including the heart),
4. Ω_H is the domain of the heart,
5. M and M_i are the conductivity tensors of the body and heart, respectively.

See Figure 1.1 for a visual representation of the domains.

Intuitively, it might seem peculiar that (1.1) is time-independent, given the fact that the potentials in the heart vary over the course of a heartbeat. It is known, however, that the transmembrane potential is approximately piecewise constant during the ST-segment of the heartcycle:

$$f(x) \approx \begin{cases} 0mV & x \text{ in healthy tissue,} \\ 50mV & x \text{ in ischemic tissue.} \end{cases} \tag{1.2}$$

This is supported by biomedical evidence, and the voltages used in (1.2) assume that the potentials have been normalized with respect to rest. For further details, see e.g. [5].

Consequently, if we can find the region of the heart Ω_H where the transmembrane potential is approximately equal to $50mV$, we can determine the ischemic region from (1.2).

In diagnostics, to measure the transmembrane potential directly is not realistic, but the extracellular potential on the body surface is readily available from ECG recordings. Thus, we obtain the inverse problem

$$\min_{(f,u)\in H^1(\Omega_H)\times H^1(\Omega_B)} \left\{ \frac{1}{2}\|Tu - d\|^2_{L^2(\partial\Omega_B)} + \frac{1}{2}\alpha\|f\|^2_{H^1(\Omega_H)} \right\} \tag{1.3}$$

subject to

$$\int_{\Omega_B} \nabla\psi \cdot M\nabla u \, dx = -\int_{\Omega_H} \nabla\psi \cdot M_i \nabla f \, dx, \ \forall\psi \in H^1(\Omega_B), \tag{1.4}$$

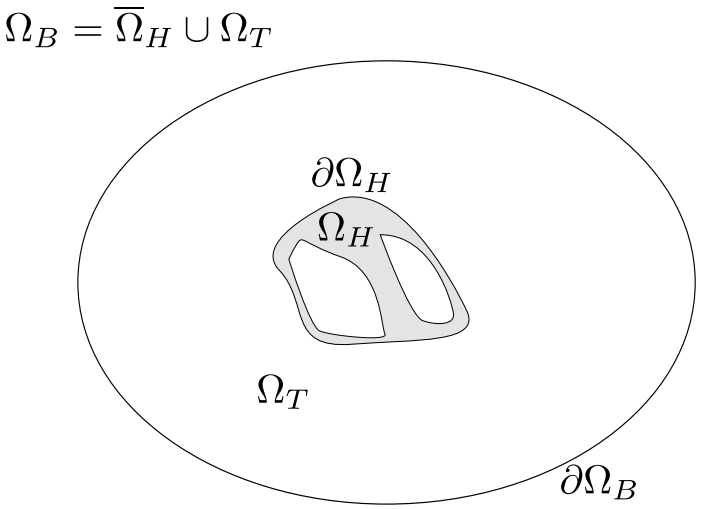


Figure 1.1: A 2D picture of the domains. Ω_H represents the heart and is depicted in gray color. We denote the remaining domain by the torso, Ω_T. The cavities (white areas) inside the heart represent the ventricles.

where $d \in L^2(\partial\Omega_B)$ is the ECG recording and $T : H^1(\Omega_B) \to L^2(\partial\Omega_B)$ is the trace operator.

If this model is to be used in clinical practice, there are several challenges which need to be addressed. All the problems studied in this thesis can, as aforementioned, be motivated from a desire to make the inverse ECG problem applicable for medical use. We present the different issues studied in this thesis, in bullet points.

1. To enhance the accuracy of the model, one might attempt to incorporate more information. We already mentioned that, according to biomedical knowledge, the transmembrane potential is known *a priori* to approximately satisfy (1.2). This motivates the additional constraint

$$0 \leq f(x) \leq 50 \quad \forall x \in \Omega_H.$$

   In Paper I, such box constraints are studied for controls in L^2, and a numerical investigation is also undertaken for an H^1-control function.

2. It is well known that Tikhonov regularization, which is applied in (1.3), yields smooth solutions. From a diagnostics perspective, however, it might be beneficial to clearly separate the ischemic and the healthy regions. The Tikhonov regularization technique will in such cases be of limited value. To allow for discontinuous solutions, we can instead

apply Total Variation (TV) regularization, i.e.

$$\mathcal{R}(f) = \alpha \int_{\Omega_H} |\nabla f| \, dx. \tag{1.5}$$

Formally, the functional $\mathcal{R}$ can also be applied in a distributional sense to functions in a weaker space than $W^{1,1}(\Omega_H)$. This alternative regularization technique is studied in Paper II.

3. In the PDE-constrained optimization community, one of the most active research topics is preconditioning of the optimality systems associated with the optimization problem. This is essential in order to obtain efficient numerical schemes. In Paper III, we study this issue for optimization problems associated with (1.3)-(1.4). Essentially, the challenge is to find the "natural" subspaces, of the general Sobolev spaces, to which the control and state functions belong.

4. Finally, we are concerned with the domains Ω_B and Ω_H associated with (1.3)-(1.4). In software, the representation of these domains is challenging. The domains are rather complex, patient specific, and a segmentation of the body from MRI data is still quite time consuming and might involve manual labor to segment blurred or unclear transitions between regions. Therefore, we are interested in domain embedding techniques, and in particular the diffuse domain method, which basically only relies upon a signed distance function. The signed distance function measure the Euclidean distance to the heart, and is less challenging to derive than performing a segmentation, and it does not involve actual meshing of the patient specific heart. In the last two papers (IV and V), we investigate how complex or unknown domains can be embedded inside larger and trivial domains, and how we can approximate and solve the optimization problem on this larger domain.

1.1.2 Linear PDE-constrained optimization problems

For optimal control problems with linear PDE constraints, we can formulate the abstract optimization problem

$$\min_{(f,u)\in F\times U} \underbrace{\left\{\frac{1}{2}\|Tu-d\|_Z^2 + \frac{1}{2}\alpha\|f\|_F^2\right\}}_{J(f,u)}, \tag{1.6}$$

subject to

$$Au + Bf = 0, \quad f \in F_{ad}, \tag{1.7}$$

where $\alpha > 0$ is the regularization parameter, d is the measured observation data, and the control function f and state function u are defined on the

domains Ω_f and Ω, i.e.

$$f : \Omega_f \to \mathbb{R}, \quad u : \Omega \to \mathbb{R}, \quad \Omega_f \subseteq \Omega. \tag{1.8}$$

Furthermore, we introduce the following assumptions:

Assumption 1.1.

1. *F, U and Z are Hilbert spaces,*
2. *$F_{ad} \subset F$ is non-empty, closed and convex,*
3. *$A : U \to U'$ is linear, bounded and has a bounded inverse*[2]*,*
4. *$B : F \to U'$ is linear and bounded, and*
5. *$T : U \to Z$ is linear and bounded.*

The symbol "$'$" is used to denote dual spaces and dual operators, i.e.

$$\langle Au, \phi \rangle = \langle A'\phi, u \rangle \quad \forall u, \phi \in U.$$

The first term in (1.6) is known as the *fidelity* term, whereas the second term is called the *regularization* term. For most PDE-constrained optimization problems, the observation data d is measured on a restricted domain, e.g. the boundary of the domain of the state equation. With such limited observations and $\Omega_f \subset\subset \Omega$, a lack of regularization, i.e. $\alpha = 0$, will typically result in (1.6)-(1.7) being severely ill-posed, and a solution might not even be unique. However, the following theorem asserts when a unique solution is guaranteed.

Theorem 1.1. *Let $\alpha > 0$ and assume that Assumption 1.1 holds. Then there exists a unique solution of* (1.6)-(1.7)*.*

Proof. A proof can be found in [3, Theorem 1.43]. □

1.1.3 First-order optimality condition

There are several solution methods for constrained optimization problems. In the field of PDE-constrained optimization, one technique that has received much attention is the "all-at-one" approach, where one solves the entire corresponding first-order optimality system simultaneously. For convex optimization problems, these optimality systems, known as the Karush-Kuhn-Tucker (KKT) conditions, yield necessary, and sometimes sufficient, criteria for an optimal solution to exist.

[2] In general $A : U \to V'$, if we apply test functions which are different from the trial functions. We do not encounter such formulations in this thesis, and hence we restrict the focus to $A : U \to U'$.

To derive the KKT conditions for the optimization problem (1.6)-(1.7), we first define the reduced functional $\hat{J} : F \to \mathbb{R}$ by

$$\hat{J}(f) = J(f, -A^{-1}Bf), \tag{1.9}$$

where $u = -A^{-1}Bf$ is the solution of (1.7) and J is defined in (1.6). Secondly, we introduce the notion of Riesz maps. For a general Hilbert space $\mathcal{H}$, the Riesz map

$$R_{\mathcal{H}} : \mathcal{H} \to \mathcal{H}'$$

is the canonical isometry between $\mathcal{H}$ and its dual space $\mathcal{H}'$. The optimality conditions then read

Theorem 1.2 (Necessary and sufficient first-order conditions)**.** *Let $\alpha > 0$ and assume that Assumption 1.1 holds. Then there exists $\bar{w} \in U$ such that $(\bar{f}, \bar{u}) \in F_{ad} \times U$ is the optimal solution of* (1.6)-(1.7) *if and only if the following conditions are satisfied:*

$$\begin{aligned} \langle \alpha R_F \bar{f}, f - \bar{f} \rangle + \langle B'\bar{w}, f - \bar{f} \rangle &\geq 0, \quad \forall f \in F_{ad}, && (1.10)\\ T'T\bar{u} + A'\bar{w} &= T'd, && (1.11)\\ A\bar{u} + B\bar{f} &= 0. && (1.12) \end{aligned}$$

Proof. A proof can be found in the standard literature, but we include one for the sake of completeness. By assumption, F_{ad} is non-empty and convex. Furthermore, $\hat{J}$ in (1.9) is strictly convex, Gâteaux differentiable and coercive, i.e.

$$\hat{J}(f) \to \infty \quad \text{if } \|f\|_F \to \infty.$$

From standard convex optimization theory, a necessary and sufficient condition for a (unique) optimal solution $\bar{f} \in F_{ad}$ of $\min_f \hat{J}(f)$ is then

$$\langle \hat{J}'(\bar{f}), f - \bar{f} \rangle \geq 0 \quad \forall f \in F_{ad}. \tag{1.13}$$

The derivative of $\hat{J}$ is

$$\begin{aligned} \hat{J}'(\bar{f}) &= \alpha R_F \bar{f} + B'[A']^{-1}T'(TA^{-1}B\bar{f} + d)\\ &= \alpha R_F \bar{f} + B'\bar{w} \in F', \end{aligned}$$

where we have defined

$$\bar{w} = [A']^{-1}T'(TA^{-1}B\bar{f} + d). \tag{1.14}$$

Substituting this into (1.13), we find that

$$\langle \hat{J}'(\bar{f}), f - \bar{f} \rangle = \langle \alpha R_F \bar{f}, f - \bar{f} \rangle + \langle B'\bar{w}, f - \bar{f} \rangle \geq 0 \quad \forall f \in F_{ad},$$

which yields (1.10). Furthermore, since A is a bijection, see Assumption 1.1, it immediately follows that

$$\bar{u} = -A^{-1}B\bar{f}, \tag{1.15}$$

which yields (1.12). Also, from the definition (1.14) of $\bar{w}$ we note that

$$A'\bar{w} = T'(TA^{-1}B\bar{f} + d) = -T'T\bar{u} + T'd,$$

see (1.15), and consequently (1.11) follows. □

Remark 1.1 (Lagrangian)**.** *If we introduce the Lagrangian* $\mathcal{L} : F \times U \times U \to \mathbb{R}$ *as*

$$\begin{aligned} \mathcal{L}(f,u,w) &= J(f,u) + \langle Au + Bf, w\rangle \\ &= \frac{1}{2}\|Tu - d\|_Z^2 + \frac{1}{2}\alpha\|f\|_F^2 + \langle Au + Bf, w\rangle, \end{aligned}$$

see (1.6)-(1.7)*, we observe that*

$$\begin{aligned} \mathcal{L}_f(f,u,w) &= \alpha R_F f + B'w, \\ \mathcal{L}_u(f,u,w) &= T'(Tu - d) + A'w, \\ \mathcal{L}_w(f,u,w) &= Au + Bf. \end{aligned}$$

Consequently, the optimality conditions in Theorem 1.2 can be formulated as

$$\begin{aligned} \langle \mathcal{L}_f(\bar{f},\bar{u},\bar{w}), f - \bar{f}\rangle &\geq 0, \forall f \in F_{ad}, \\ \mathcal{L}_u(\bar{f},\bar{u},\bar{w}) &= 0, \\ \mathcal{L}_w(\bar{f},\bar{u},\bar{w}) &= 0. \end{aligned} \tag{1.16}$$

Hence, the function w*, introduced in Theorem 1.2, is called a Lagrange multiplier.*

Remark 1.2 (Optimality system without control constraints)**.** *If* $F_{ad} = F$*, the condition* (1.16) *in Remark 1.1 becomes*

$$\mathcal{L}_f(\bar{f},\bar{u},\bar{w}) = 0.$$

1.2 Preconditioners

In the first three papers, we study optimality systems which need efficient iterative solvers. Although these systems are not identical, they carry a similar structure.

To discuss this structure, let us first consider the case $F_{ad} = F$. Recall from remarks 1.1 and 1.2 that the optimality conditions for (1.6)-(1.7) then read

$$\begin{aligned} \mathcal{L}_f(\bar{f},\bar{u},\bar{w}) &= 0, \\ \mathcal{L}_u(\bar{f},\bar{u},\bar{w}) &= 0, \\ \mathcal{L}_w(\bar{f},\bar{u},\bar{w}) &= 0, \end{aligned}$$

or

$$\begin{aligned} \alpha R_F f + B'w &= 0, && (1.17)\\ T'(Tu-d) + A'w &= 0, && (1.18)\\ Au + Bf &= 0. && (1.19) \end{aligned}$$

In order to study the optimality conditions as a so-called saddle-point system, we introduce the two bilinear forms $a : (F \times U) \times (F \times U) \to \mathbb{R}$ and $b : (F \times U) \times U \to \mathbb{R}$ by

$$\begin{aligned} a(f,u;\phi,\psi) &= \alpha\langle R_F f, \phi\rangle + \langle T'Tu, \psi\rangle,\\ b(f,u;\varphi) &= \langle Au + Bf, \varphi\rangle. \end{aligned}$$

These bilinear forms enable us to formulate the optimality conditions in Theorem 1.2, for the case $F_{ad} = F$, as the saddle-point problem

$$\begin{aligned} a(f,u;\phi,\psi) + b(\phi,\psi;w) &= g(\phi,\psi) \quad \forall (\phi,\psi) \in F \times U,\\ b(f,u;\varphi) &= h(\varphi) \qquad \forall \varphi \in U, \end{aligned}$$

where, in our case, the functionals $g : F \times U \to \mathbb{R}$ and $h : U \to \mathbb{R}$ are defined as

$$\begin{aligned} g(\phi,\psi) &= \langle T'd, \psi\rangle,\\ h(\varphi) &= 0, \end{aligned}$$

see (1.17)-(1.19). Alternatively, (1.17)-(1.19) can be formulated on the block form

$$\underbrace{\begin{bmatrix} \alpha R_F & 0 & B' \\ 0 & T'T & A' \\ B & A & 0 \end{bmatrix}}_{\mathcal{A}_\alpha} \underbrace{\begin{bmatrix} f \\ u \\ w \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 0 \\ T'd \\ 0 \end{bmatrix}}_{q}, \qquad (1.20)$$

where

$$\mathcal{A}_\alpha : F \times U \times U \to F' \times U' \times U'.$$

For such saddle-point systems, we can apply Brezzi's splitting theorem to obtain stability estimates.

The aim of papers I-III is to create preconditioners such that the solvers are robust with respect to both the mesh parameter h and the regularization parameter α. In more abstract terms, to achieve full robustness we need weighted Sobolev spaces in which the stability estimates are independent of the parameters h and α. This is far from trivial, but if $\|\mathcal{A}_\alpha\|$ and $\|\mathcal{A}_\alpha^{-1}\|$ are bounded independently of h, solving (1.20) with h-independent iteration numbers can be achieved by applying an appropriate Krylov solver to the preconditioned system

$$\mathcal{B}_\alpha \mathcal{A}_\alpha x = \mathcal{B}_\alpha q, \qquad (1.21)$$

where $\mathcal{B}_\alpha : F' \times U' \times U' \to F \times U \times U$ is an isomorphism, with h-independent bounds for both $\|\mathcal{B}_\alpha\|$ and $\|\mathcal{B}_\alpha^{-1}\|$, as discussed in [4]. With a sound discretization of (1.21), the h-independence is inherited by the associated discretized system.

Concerning the regularization parameter, complete α-robust solvers for (1.20) have only been obtained for a few specific state equations where $\Omega_f = \Omega$ and the data is measured on the entire domain Ω, see [10, 8]. If $\Omega_f \subset\subset \Omega$ and only limited observation data is available, the most general result is a maximum growth in iteration numbers of order $O([\log(\alpha^{-1})]^2)$ as $\alpha \to 0$, see [6]. The authors of [6] also explain why growth of order $O(\log(\alpha^{-1}))$ often is to be expected in practice.

We now briefly discuss the optimality systems we study in papers I-III, and what degree of α-robustness we obtain in each case:

1. In the first paper, we invoke the so-called Primal Dual Active Set (PDAS) algorithm to deal with the box constrains

$$f_l(x) \leq f(x) \leq f_u(x) \quad \forall x \in \Omega_f,$$

   on the control. (See Section 1.3.1 for a brief introduction to the PDAS method.) This results in an iterative procedure where we solve a sequence of equations of the form (1.20). In each iteration, the control f is only unknown on parts of the domain, so the mappings in the first column of $\mathcal{A}_\alpha$ operates on functions with restricted support. We prove a maximum growth of order $O([\log(\alpha^{-1})]^2)$ in iteration numbers when $\alpha \to 0$, and present experiments with logarithmic growth.

2. The second paper is concerned with total variation (TV) regularization. We apply the split Bregman algorithm to deal with the non-differentiability of the regularization term, see (1.5). (For an overview of the split Bregman method, see Section 1.3.2.) In each iteration, we must solve a system similar to (1.20), only with R_F replaced by a weak form of $-\Delta$. Also in this paper we prove a maximum growth in iteration numbers of order $O([\log(\alpha^{-1})]^2)$.

3. In Paper III we have no box constraints and a standard Tikhonov regularization term. For a specific control space and state equation, we are able to obtain full α-robustness in the case of $\Omega_f \subset\subset \Omega$, cf. (1.8), and observation data only on the boundary $\partial\Omega$ of Ω. In more detail, if $F = H^1(\Omega_f)'$ and $U = H^1(\Omega)$, it can be shown that the state function and Lagrange multiplier actually belong to a subspace of $H^1(\Omega)$ which is isomorphic to the control space $H^1(\Omega_f)'$, leading to stability estimates independent of α.

1.3 Algorithms

Two of the algorithms applied in this thesis are not standard techniques in optimization, and we will therefore sketch the ideas behind the *Primal Dual Active Set* (PDAS) and *(split) Bregman* methods.

1.3.1 PDAS method

In Paper I we investigate optimization problems with box constraints on the control. To show how the PDAS method can be used to solve such problems, consider the optimization problem

$$\min_{f\in L^2(\Omega_f)} \hat{J}(f) \quad \text{s.t.} \quad f(x) \geq 0. \tag{1.22}$$

We only consider the box constraint $f(x) \geq 0$ to avoid unnecessary details.

If $\bar{f}$ is an optimal solution of (1.22), we can define the active set $\bar{\mathcal{A}}$ and the inactive set $\bar{\mathcal{I}}$ by

$$\bar{\mathcal{A}} = \{x \in \Omega_f : \bar{f}(x) = 0\} \quad \text{and} \quad \bar{\mathcal{I}} = \bar{\mathcal{A}} \setminus \Omega_f. \tag{1.23}$$

If we define the Lagrange functional

$$\mathcal{L}(f, \lambda) = \hat{J}(f) - (f, \lambda)_{L^2(\Omega_f)},$$

associated with (1.22), we obtain the well-known first-order conditions

$$\mathcal{L}_f(\bar{f}, \bar{\lambda}) = \hat{J}'(\bar{f}) - \bar{\lambda} = 0, \tag{1.24}$$

$$\bar{\lambda} \geq 0, \quad \bar{f} \geq 0, \quad \bar{\lambda}\bar{f} = 0, \tag{1.25}$$

see e.g. [7].

Since the active set $\bar{\mathcal{A}}$ in (1.23) is unknown, we need an algorithm to find this set. Due to the condition (1.25), the most intuitive approach is maybe to guess an active set $\mathcal{A}^0$, construct the inactive set $\mathcal{I}^0 = \Omega_f \setminus \mathcal{A}^0$, and then solve

$$\begin{aligned} \hat{J}'(f^{(1)}) &= \lambda^{(1)}, \\ f^{(1)} &= 0, \quad x \in \mathcal{A}^0, \\ \lambda^{(1)} &= 0, \quad x \in \mathcal{I}^0. \end{aligned}$$

Next, we update our inactive and active sets in accordance with whether (1.25) is satisfied. That is,

1. If $x \in \mathcal{A}^0$ and $\lambda^{(1)}(x) \leq 0$, we move x to the inactive set $\mathcal{I}^1$.
2. If $x \in \mathcal{I}^0$ and $f^{(1)}(x) < 0$, we move x to the active set $\mathcal{A}^1$.
3. Otherwise, (1.25) holds, and x stays in the same set.

After updating the sets $\mathcal{A}^1$ and $\mathcal{I}^1$, we can solve

$$\begin{aligned} \hat{J}'(f^{(2)}) &= \lambda^{(2)}, \\ f^{(2)} &= 0, \quad x \in \mathcal{A}^1, \\ \lambda^{(2)} &= 0, \quad x \in \mathcal{I}^1, \end{aligned}$$

etc.

We keep doing this in an iterative manner until convergence is reached. This is essentially the PDAS algorithm.

To formulate the algorithm more precisely, we start by observing that (1.25) is equivalent to the condition

$$\forall c > 0: \quad \lambda(x) + \min(0, cf(x) - \lambda(x)) = 0. \tag{1.26}$$

From this observation, we can define the active and inactive sets depending on whether a point x violates (1.26), i.e. whether $\lambda(x) + \min(0, cf(x) - \lambda(x)) \neq 0 \quad \forall c > 0$. The two sets become

$$\begin{aligned} \mathcal{A}^k &= \{x \in \Omega_f : (cf^k - \lambda^k)(x) < 0\}, \\ \mathcal{I}^k &= \Omega_f \setminus \mathcal{A}^k. \end{aligned}$$

By combining these updates with solving

$$\begin{aligned} \hat{J}'(f^{k+1}) &= \lambda^{k+1}, \\ f^{k+1} &= 0, \quad x \in \mathcal{A}^k, \\ \lambda^{k+1} &= 0, \quad x \in \mathcal{I}^k, \end{aligned}$$

we obtain the full algorithm. In more rigorous terms, the method is a specific case of a semismooth Newton method. We will not go into details, but refer to [2, 3].

1.3.2 Bregman and split Bregman methods

Our motivation for introducing the split Bregman method is its success in solving finite dimensional optimization problems of the form

$$\min_{f_h \in F_h} \left\{ \frac{1}{2} \|K_h f_h - d_h\|_{Z_h}^2 + \alpha \int_{\Omega_f} |\nabla f_h| \right\}, \tag{1.27}$$

i.e. problems with TV regularization. We can relate (1.27) to a PDE-constrained optimization problem if K_h is a discrete approximation of the operator $K : F \to Z$, defined as

$$K = -TA^{-1}B,$$

see (1.6)-(1.7).

The split Bregman method is derived from the Bregman method, which again can be understood as a generalization of the classical proximal point method. Hence, we start by briefly presenting the latter method.

Proximal method

Consider the general optimization problem

$$\min_{x\in\mathcal{H}_1} \mathcal{F}(x), \tag{1.28}$$

where $\mathcal{H}_1$ is a finite dimensional Hilbert space. The proximal point method is the iterative algorithm

$$x^{k+1} = \arg\min_{x\in\mathcal{H}_1} \left\{ \frac{1}{2\lambda}\|x - x^k\|^2_{\mathcal{H}_1} + \mathcal{F}(x) \right\}. \tag{1.29}$$

To understand the motivation behind this method, we define the proximal operator

$$prox_{\mathcal{F}}(y) = \arg\min_{x\in\mathcal{H}_1} \left\{ \frac{1}{2\lambda}\|x - y\|^2_{\mathcal{H}_1} + \mathcal{F}(x) \right\}, \tag{1.30}$$

and observe that the solution of (1.28) is a fixed point of $prox_{\mathcal{F}}(y)$. Hence, if the original minimization problem (1.28) is difficult to solve, the proximal point method allows us to solve a sequence of "nicer" problems.

Bregman method

The Bregman method can be viewed as a generalization of (1.29). For a convex and Gâteaux differentiable function $\Phi : \mathcal{H}_1 \to \mathbb{R}$, we define the *Bregman distance*

$$B_\Phi(x, y) = \Phi(x) - \Phi(y) - (\nabla\Phi(y), x - y)_{\mathcal{H}_1}, \tag{1.31}$$

where we would like to emphasize that $\nabla\Phi(y) \in \mathcal{H}_1$ is the "Riesz derivative" of $\Phi(y)$, i.e.

$$\nabla\Phi(y) = R^{-1}_{\mathcal{H}_1}\Phi'(y).$$

Since the function Φ is convex, the Bregman distance will always be positive. Following the recipe from the proximal point method, we derive the algorithm

$$x^{k+1} = \arg\min_{x\in\mathcal{H}_1} \left\{ B_\Phi(x, x^k) + \lambda\mathcal{F}(x) \right\}. \tag{1.32}$$

If we choose $\Phi(\cdot) = \frac{1}{2}\|\cdot\|^2_{\mathcal{H}_1}$, the Bregman algorithm reduces to the standard proximal point method.

The main strength of the Bregman method, however, is that the function Φ is not required to be differentiable. For convex functions, the derivative $\nabla\Phi(\cdot)$ is generalized by the set-valued subdifferential $\partial\Phi(\cdot)$ of $\Phi(\cdot)$.

If the derivative does not exist, we must choose a specific element, a subderivative p, in the set-valued subdifferential. To explain how to consistently select such a subderivative, assume that $p^k \in \partial\Phi(x^k)$. Then, replacing

$\nabla\Phi(x^k)$ by p^k in (1.31)-(1.32) yields

$$\begin{aligned} x^{k+1} &= \arg\min_{x\in\mathcal{H}_1}\left\{B_\Phi^{p^k}(x,x^k)+\lambda\mathcal{F}(x)\right\} \\ &= \arg\min_{x\in\mathcal{H}_1}\left\{\Phi(x)-\Phi(x^k)-(p^k,x-x^k)_{\mathcal{H}_1}+\lambda\mathcal{F}(x)\right\} \\ &= \arg\min_{x\in\mathcal{H}_1}\left\{\Phi(x)-(p^k,x)_{\mathcal{H}_1}+\lambda\mathcal{F}(x)\right\}, \end{aligned} \tag{1.33}$$

where we have used the fact that $\Phi(x^k)$ and (p^k,x^k) are independent of x. Since x^{k+1} is a minimizer of (1.33), we know from standard optimization theory that

$$\partial\Phi(x^{k+1})-p^k+\lambda\nabla\mathcal{F}(x^{k+1})\ni 0.$$

Hence, we can choose the update

$$p^{k+1}=p^k-\lambda\nabla\mathcal{F}(x^{k+1})\in\partial\Phi(x^{k+1}), \tag{1.34}$$

to get a consistent choice for p^{k+1}.

To initialize the algorithm, it is standard to choose $x^0=0$ and $p^0=0$. For a different choice of x^0, the process of choosing p^0 becomes an optimization problem in itself. To summarize, the Bregman method consists of the two updates

$$x^{k+1} = \arg\min_{x\in\mathcal{H}_1}\left\{\Phi(x)-(p^k,x)_{\mathcal{H}_1}+\lambda\mathcal{F}(x)\right\}, \tag{1.35}$$

$$p^{k+1} = p^k-\lambda\nabla\mathcal{F}(x^{k+1}). \tag{1.36}$$

Quadratic problems

Before we address how the Bregman method can be applied to solve (1.27), we consider the quadratic minimization problem

$$\min_{x\in\mathcal{H}_1}\underbrace{\frac{1}{2}\|Lx-z\|_{\mathcal{H}_2}^2}_{\mathcal{F}(x)}, \tag{1.37}$$

where $L:\mathcal{H}_1\to\mathcal{H}_2$ is a linear operator between two finite dimensional Hilbert spaces. The derivative of $\mathcal{F}(x)$, defined in (1.37), is

$$\nabla\mathcal{F}(x)=L^*(Lx-z),$$

where $L^*:\mathcal{H}_2\to\mathcal{H}_1$ is the adjoint of L. Consequently, if we choose $p^0=0$, the update (1.36) for p^{k+1} becomes

$$p^{k+1}=-\lambda L^*\left(\sum_{n=1}^{k+1}Lx^n-z\right).$$

From this update of p^{k+1}, we can rewrite (1.35) on the form

$$x^{k+1} = \arg\min_{x\in\mathcal{H}_1}\left\{\Phi(x) + \frac{1}{2}\lambda\|Lx - z + \sum_{n=1}^{k}(Lx^n - z)\|^2_{\mathcal{H}_2}\right\}.$$

Then, by defining $b^k = \sum_{n=1}^{k}(Lx^n - z)$, we can in this case simplify (1.35)-(1.36) to

$$\begin{cases} x^{k+1} = \arg\min_x\left\{\Phi(x) + \frac{1}{2}\lambda\|Lx - z + b^k\|^2_{\mathcal{H}_2}\right\} \\ b^{k+1} = b^k + (Lx^{k+1} - z). \end{cases} \tag{1.38}$$

Split Bregman method

In [1], the authors realized that (1.38) could be applied very carefully in order to solve (1.27). This minimization problem can be formulated equivalently as the constrained problem

$$\min_{f_h,q_h}\left\{\frac{1}{2}\|K_h f_h - d_h\|^2_{Z_h} + \alpha\int_{\Omega_f}|q_h|\right\} \quad \text{s.t} \quad \nabla f_h = q_h.$$

Now, let $x = (f_h, q_h)$ and define

$$\Phi(x) = \frac{1}{2}\|K_h f_h - d_h\|^2_{Z_h} + \alpha\int_{\Omega_f}|q_h|, \tag{1.39}$$

$$\mathcal{F}(x) = \frac{1}{2}\|\nabla f_h - q_h\|^2_{L^2_h(\Omega_f)}. \tag{1.40}$$

Applying (1.38) to this choice of Φ and $\mathcal{F}$ yields the algorithm

$$\begin{cases} (f_h^{k+1}, q_h^{k+1}) & = \arg\min_{f_h,q_h}\{\frac{1}{2}\|K_h f_h - d_h\|^2_{Z_h} \\ & \quad +\alpha\int_{\Omega_f}|q_h| + \frac{1}{2}\lambda\|\nabla f_h - q_h + b^k\|^2_{L^2_h(\Omega_f)}\} \\ b^{k+1} & = b^k + (\nabla f_h^{k+1} - q_h^{k+1}). \end{cases} \tag{1.41}$$

We observe that $\min_x \mathcal{F}(x) = 0$ if $\nabla f_h = q_h$, see (1.40) and recall also the discussion of (1.28)-(1.30). Hence, if the algorithm converges to a minimizer $\bar{x} = (\bar{f}_h, \bar{q}_h)$ of $\mathcal{F}$, i.e. $\nabla\bar{f}_h = \bar{q}_h$, we observe from (1.41) that this $\bar{x}$ also minimize $\Phi(x)$, see (1.39). For more details on convergence and equivalence to other methods, see e.g. [11].

To simplify (1.41), one can split the first step of (1.41) into two minimization problems. That is, we first freeze $q_h = q_h^k$ and minimize for f_h^{k+1}, and then freeze $f_h = f_h^{k+1}$ and minimize for q_h^{k+1}. The result is the split Bregman algorithm

$$\begin{cases} f_h^{k+1} = \arg\min_{f_h}\left\{\frac{1}{2}\|K_h f_h - d_h\|^2_{Z_h} + \frac{1}{2}\lambda\|\nabla f_h - q_h^k + b^k\|^2_{L^2_h(\Omega_f)}\right\} \\ q_h^{k+1} = \arg\min_{q_h}\left\{\alpha\int_{\Omega_f}|q_h| + \frac{1}{2}\lambda\|\nabla f_h^{k+1} - q_h + b^k\|^2_{L^2_h(\Omega_f)}\right\} \\ b^{k+1} = b^k + (\nabla f_h^{k+1} - q_h^{k+1}). \end{cases} \tag{1.42}$$

Both (1.41) and (1.42) are commonly referred to as the split Bregman method, although they are not equivalent. The former is equivalent to an augmented Lagrangian method, whereas the latter is equivalent to an Alternating Direction Method of Multipliers (ADMM) algorithm, see e.g. [11] for a summary. For simulations, (1.42) is usually the preferred method.

1.3.3 Diffuse domain method

In order to solve PDEs or PDE-constrained optimization problems numerically, we need a mesh representation of the domain. This is not necessarily trivial, if the domain is complex, or even unknown. For example, meshing the heart of a patient is difficult, and requires proper segmentation. Several methods have been suggested to remedy this issue, among them the diffuse domain method.

This method relies on the fact that the signed distance function

$$d_\Omega(x) = \operatorname{dist}(x, \Omega) - \operatorname{dist}(x, \mathbb{R}^n \setminus \Omega)$$

can be used to describe the domain Ω. That is, we have $\Omega = \{x \in \mathbb{R}^n : d_\Omega(x) < 0\}$. Consequently, we can in a controllable manner embed the complex domain Ω in a larger, easily implementable domain $\mathbf{\Omega}$. See Figure 1.2 for a visual representation of the embedding.

The diffuse domain method then makes use of the distance function d_Ω in order to approximate second order elliptic boundary value problems posed on Ω by variational forms given on the larger domain $\mathbf{\Omega}$. To illustrate this procedure in detail, let us consider the following PDE

$$\begin{aligned} -\nabla \cdot (M \nabla u) + cu &= f \quad \text{in } \Omega, \\ \mathbf{n} \cdot M \nabla u + bu &= g \quad \text{on } \partial\Omega. \end{aligned}$$

The weak formulation of this boundary value problem reads: Find u such that for all suitable test functions v,

$$\int_\Omega (\nabla v \cdot M \nabla u + cuv)\, dx + \int_{\partial\Omega} buv\, d\sigma = \int_\Omega fv\, dx + \int_{\partial\Omega} gv\, d\sigma. \tag{1.43}$$

Written more generically, we have expressions of the form

$$\int_\Omega k(x) dx \quad \text{and} \quad \int_{\partial\Omega} l(x) d\sigma, \tag{1.44}$$

where k and l involve test and trial functions or test and source functions.

The idea is now to approximate the integrals in (1.44) with integrals over the larger domain $\mathbf{\Omega}$. Recall that the signed distance function is negative for points inside Ω and positive for points outside Ω. Thus, if we introduce

$$\varphi^\epsilon(x) = S\left(-\frac{d_\Omega(x)}{\epsilon}\right),$$

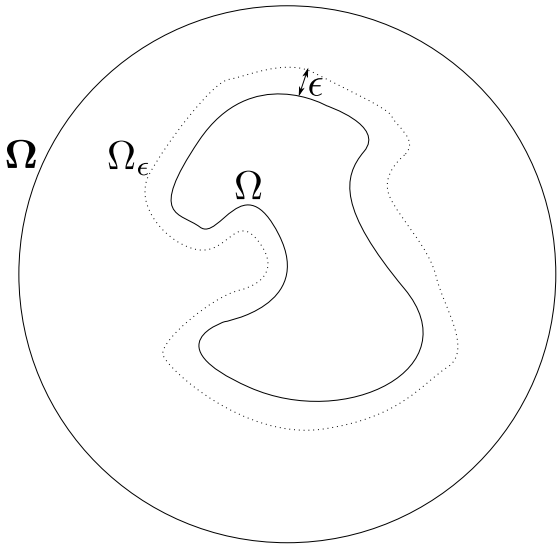


Figure 1.2: An example of a complex domain Ω, the extended domain $\Omega_\epsilon = \{x \in \mathbb{R}^n : d_\Omega(x) < \epsilon\}$ and the simple, larger domain $\boldsymbol{\Omega}$.

where S is a sigmoid function with $S(t) = \frac{t}{|t|}$ when $|t| \geq 1$, we observe that $\varphi^\epsilon(x) > 0$ for $x \in \Omega$ and $\varphi^\epsilon(x) < 0$ for $x \in \boldsymbol{\Omega} \setminus \overline{\Omega}$. Furthermore, $S(\cdot/\epsilon)$ converges to the sign function as $\epsilon \to 0$, and consequently the phase-field function

$$\omega^\epsilon = \frac{1}{2}(1 + \varphi^\epsilon). \tag{1.45}$$

converges to the indicator function for Ω as $\epsilon \to 0$.

In Paper IV, we use Fubini's theorem and the co-area formula to derive just approximations of the integrals in (1.44). Here, we will give a more intuitive justification for the approximations.

Since the phase-field function ω^ϵ is an approximation of the indicator function for Ω, it seems reasonable that

$$\int_\Omega k(x)\ dx \approx \int_{\boldsymbol{\Omega}} k(x)\omega^\epsilon(x)\ dx$$

for small values of $\epsilon > 0$.

Similarly, the absolute value $|\nabla\omega^\epsilon|$ of the gradient of ω^ϵ becomes a concentrated distribution around $\{x : d_D(x) = 0\}$, i.e. around the boundary $\partial\Omega$. Therefore, it seems reasonable that

$$\int_{\partial\Omega} l(x)\ dx \approx \int_{\boldsymbol{\Omega}} l(x)|\nabla\omega^\epsilon(x)|\ dx$$

for small values of $\epsilon > 0$.

If we apply these approximations, the original weak PDE (1.43) suggests the following variational problem: Find u^ϵ such that for all suitable test functions v,

$$\int_{\mathbf{\Omega}} (\nabla v \cdot M \nabla u^\epsilon + c u^\epsilon v)\omega^\epsilon + \int_{\mathbf{\Omega}} b u^\epsilon v |\nabla \omega^\epsilon| = \int_{\mathbf{\Omega}} f v \omega^\epsilon + \int_{\mathbf{\Omega}} g v |\nabla \omega^\epsilon|, \quad (1.46)$$

where M, f and g are extended carefully to the larger domain $\mathbf{\Omega}$.

In Paper IV, convergence rates of

$$\|u^\epsilon - u\|_{\mathcal{X}}$$

is studied in different norms and under different regularization assumptions.

In Paper V, the diffuse domain formulation is applied to a PDE-constrained optimization problem. That is, all original variational forms in the associated KKT system are approximated with diffuse variational forms similar to (1.46). Here, we are also interested in robust preconditioners of the diffuse KKT systems - not only with respect to the regularization parameter α and the mesh parameter h, but also with respect to the diffuse parameter ϵ. We derive preconditioners with complete ϵ-robustness.

Bibliography

[1] T. Goldstein and S. Osher. The split Bregman method for L1-regularized problems. *SIAM Journal on Imaging Sciences*, 2:323–343, 2009.

[2] M. Hintermüller, K. Ito, and K. Kunisch. The Primal-Dual Active Set strategy as a semismooth Newton method. *SIAM Journal on Optimization*, 13(3):865–888, 2003.

[3] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints*. Springer-Verlag, 2009.

[4] K. A. Mardal and R. Winther. Preconditioning discretizations of systems of partial differential equations. *Numerical Linear Algebra with Applications*, 18(1):1–40, 2011.

[5] B. F. Nielsen, M. Lysaker, and P. Grøttum. Computing ischemic regions in the heart with the bidomain model - First steps towards validation. *IEEE Transactions on Medical Imaging*, 32:1085–1096, 2013.

[6] B. F. Nielsen and K. A. Mardal. Analysis of the Minimal Residual Method applied to ill-posed optimality systems. *SIAM Journal on Scientific Computing*, 35(2):A785–A814, 2013.

[7] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2 edition, 2006.

[8] J. W. Pearson and A. J. Wathen. A new approximation of the Schur complement in preconditioners for PDE-constrained optimization. *Numerical Linear Algebra with Applications*, 19:816–829, 2012.

[9] A. J. Pullan, M. L. Buist, and L. K. Cheng. *Mathematically Modelling the Electrical Activity of the Heart: From Cell to Body Surface and Back*. World Scientific Publishing Company, 2005.

[10] J. Schöberl and W. Zulehner. Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems. *SIAM Journal on Matrix Analysis and Applications*, 29(3):752–773, 2007.

[11] S. Setzer. Operator splitting, Bregman methods and frame shrinkage in image processing. *International Journal of Computer Vision*, 92:265–280, 2011.

[12] J. Sundnes, G. T. Lines, X. Cai, B. F. Nielsen, K. A. Mardal, and A. Tveito. *Computing the Electrical Activity in the Heart*. Springer-Verlag, 2006.

2

Paper I - Preconditioners for PDE-constrained optimization problems with box constraints: Towards high resolution inverse ECG images

This paper is submitted for publication.

Preconditioners for PDE-constrained optimization problems with box constraints: Towards high resolution inverse ECG images

Ole Løseth Elvetun*and Bjørn Fredrik Nielsen†

June 25, 2015

Abstract

By combining the Minimal Residual Method and the Primal-Dual Active Set algorithm, we derive an efficient scheme for solving a class of PDE-constrained optimization problems with inequality constraints. The approach studied in this paper addresses box constrains on the control function, and leads to an iterative scheme in which linear optimality systems must be solved in each iteration. We prove that the spectra of the associate saddle point operators, appearing in each iteration, are well behaved: Almost all the eigenvalues are contained in three bounded intervals, not containing zero. In fact, for severely ill-posed problems, the number of eigenvalues outside these three intervals are of order $O(\ln(\alpha^{-1}))$ as $\alpha \to 0$, where α is the parameter employed in the Tikhonov regularization. Krylov subspace methods are well known to handle such systems of algebraic equations very well, and we thus obtain a fast method for PDE-constrained optimization problems with box constraints. In contrast to previous papers, our investigation is not targeted at analyzing a specific model, but instead covers a rather large class of problems.

Our theoretical findings are illuminated by several numerical experiments. An example covered by our theoretical findings, as well as cases not fulfilling all the assumptions needed in the analysis, are presented. Also, in addition to computations only involving synthetic data, we briefly explore whether these new techniques can be applied to real world problems. More specifically, the algorithm is tested on a medical imaging problem with clinical patient data. These tests suggest that the method is fast and reliable.

*Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, Norway. Email: ole.elvetun@nmbu.no

†Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, Norway; Simula Research Laboratory; Center for Cardiological Innovation, Oslo University Hospital. Email: bjorn.f.nielsen@nmbu.no

Keywords: PDE-constrained optimization, Primal-Dual Active Set, Minimal Residual Method, Real World Applications.

AMS subject classifications: 65F22, 49J20, 35Q93, 65K10

1 Introduction

In the field of optimization many researchers have studied the minimization of quadratic cost-functionals with constraints given by partial differential equations. Several books have been written about this subject, see e.g [3, 5, 7, 15]. By using the Lagrange multiplier technique, one might derive a system of equations which must be satisfied by the optimal solution. After suitable discretization, this system, which typically is a saddle-point problem, can be solved by an all-at-once method. That is, a scheme in which the primal, dual and optimality conditions are solved in a fully coupled manner.

Such optimality systems are often ill-posed, which leads to bad condition numbers for the discretized systems, and regularization techniques must therefore be invoked. Typically, if Tikhonov regularization is employed, then the spectral condition number of the system is of order $O(\alpha^{-1})$, where $\alpha > 0$ is the regularization parameter. Hence one might expect that, for small values of α, the number of iterations required to solve the system, using e.g. Krylov subspace methods, would be large. However, in [11] the authors prove that the spectrum of the optimality system consists of three bounded intervals and a very limited number of isolated eigenvalues outside these three intervals. This result is established for a quite broad class of PDE constrained optimization problems and imply that the Minimal Residual Method (MINRES) will handle the associated algebraic systems very well. In fact, if the problem at hand is severely ill-posed, then the required number of iterations cannot grow faster than $O([\ln(\alpha^{-1})]^2)$ as $\alpha \to 0$, and in practice one often observes iterations counts of order $O(\ln(\alpha^{-1}))$.

Many real world problems are not only modeled by PDEs, but also involve inequality constraints. These are often given in the form of box constraints on the control function. In this paper we explore whether the method and analysis presented in [11] can be extended to handle such problems adequately.

Inequality constraints typically require the use of an iterative method to solve the overall optimization task. In consequence, since the linear systems arising in each iteration typically are ill-posed, we need to solve a sequence of algebraic systems with bad condition numbers.

For some specific state equations, such problems have been solved efficiently, see e.g. [4, 14]. These efficient techniques also combines the cherished PDAS method in [2] with different numerical techniques for solving saddle-point problems [1]. We will consider such optimization tasks in a

more abstract and general setting. More precisely, our analysis concerns the class of problems that can be written on the form

$$\min_{(v,u)\in L^2(\Omega_v)\times U} \left\{ \frac{1}{2}\|Tu-d\|_Z^2 + \frac{1}{2}\alpha\|v\|_{L^2(\Omega_v)}^2 \right\}, \tag{1}$$

subject to

$$Au + Bv = 0, \tag{2}$$
$$v(x) \geq 0 \text{ a.e. in } \Omega_v, \tag{3}$$

where

- $L^2(\Omega_v)$ is the control space,
- U is the state space, $1 \leq \dim(U) \leq \infty$, and
- Z is the observation space, $1 \leq \dim(Z) \leq \infty$.

We assume that U and Z are Hilbert spaces. Further, $\Omega_v \subset \mathbb{R}^n$ is the domain the control function v is defined on, d is the given observation data, and $\alpha > 0$ is the regularization parameter. In Section 2 we will state the assumptions we need on the linear operators A, B and T. Also, there exists a solution to the problem (1)-(3) under fairly loose assumptions. For $\alpha > 0$, the solution is unique, see e.g. [5] for details.

For the problem (1)-(2), without the inequality constraint $v(x) \geq 0$, it was proven in [11] that for a sound discretization of the associated KKT system

$$\underbrace{\begin{bmatrix} \alpha I & 0 & B^* \\ 0 & T^*T & A^* \\ B & A & 0 \end{bmatrix}}_{=\mathcal{B}_\alpha} \begin{bmatrix} v \\ u \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ T^*d \\ 0 \end{bmatrix}, \tag{4}$$

the eigenvalues of the discretized operator $\mathcal{B}_\alpha^h$ satisfies

$$\text{sp}(\mathcal{B}_\alpha^h) \subset [-b,-a] \cup [c\alpha, 2\alpha] \cup \{\lambda_1, \lambda_2, ..., \lambda_{N(\alpha)}\} \cup [a,b]. \tag{5}$$

Here, a, b and c are constants, independent of the regularization parameter α, and $N(\alpha) = O(\ln(\alpha^{-1}))$ for severely ill-posed problems. Krylov subspace methods handle problems with spectra on the form (5) very well, and, since we have an indefinite system, the Minimal Residual (MINRES) method [12] is well suited for solving (4).

Based on this discussion, we can formulate the objectives of this paper as follows:

- We will combine the PDAS method, presented in [2], with the MINRES method used in [11] to obtain a standard recipe for solving problems

of the form (1)-(3). We prove that in each iteration of the PDAS algorithm we obtain a reduced system with a spectrum on the form (5), which we then can solve efficiently with the MINRES algorithm. Our derivation of the reduced systems, arising in the PDAS method, is heavily inspired by [4, 14]. Moreover, in the numerical experiments section, we show how to apply Riesz maps as preconditioners to solve some model problems.

- Real world problems often involve highly unstructured meshes and noisy data. Our second objective is to undertake a numerical investigation of such a real world PDE-constrained optimization problem, known as *the inverse problem of electrocardiography (ECG)*. The aim is to identify a heart infarct using ECG recordings and PDE-constrained optimization with box constraints. This problem has an H^1-control function, and is therefore not supported by the analysis of (1)-(3). Nevertheless, our scheme converged, and seemed to improve the quality of the solution - compared to the solution without box constraints.

For practical PDE-constrained optimization problems, the condition numbers of the discretized KKT systems is known to increase significantly, not only as the regularization parameter $\alpha \to 0$, but also when the mesh parameter $h > 0$ decreases. We will not discuss this generally, but for the synthetic model problem, we will explain how to handle the h-dependency by invoking Riesz maps as multigrid preconditioners. We then obtain an algorithm robust with respect to h and which grows moderately in iteration numbers as $\alpha \to 0$.

Remark 1.1. *We consider the prototypical inequality constraint $v(x) \geq 0$, since the aim of this paper is to show that the linear systems occurring in each iteration of the PDAS algorithm can be efficiently solved with the MINRES method, and the simple constraint $v(x) \geq 0$ makes the derivation and analysis more transparent. To see how to handle the more general box constraints*

$$v_l(x) \leq v(x) \leq v_u(x),$$

see e.g. [16, 14]. Also note that the requirement $v(x) \geq 0$ occurs in many applications, e.g., when the control function v measures density, temperature, mass or pressure.

2 Assumptions

We assume that:

$\mathcal{A}\mathbf{1}$: $A : U \to U$ is bounded and linear[1]

[1]Assume that the state equation (2) is a PDE. Then, A is typically a mapping from

$\mathcal{A}\mathbf{2}$: A^{-1} exists and is bounded.

$\mathcal{A}\mathbf{3}$: $B : L^2(\Omega_v) \to U$ is bounded and linear.

$\mathcal{A}\mathbf{4}$: $T : U \to Z$ is bounded and linear.

$\mathcal{A}\mathbf{5}$: The optimization problem (1)-(2) is severely ill-posed for $\alpha = 0$.

As shown in [11], if the assumptions listed above hold, then for a sound discretization of the KKT system (4), the eigenvalues of this discretized system satisfies (5). If (4) is well posed for $\alpha = 0$, then the numerical solution of this problem is "straightforward" and regularization is not needed. We will focus on the challenging case, i.e. severely ill-posed systems.

3 KKT system

We will now derive the algorithm for solving (1)-(3). The first thing we need, is the optimality system, which can be obtained from the Lagrangian

$$\mathcal{L}(v,u,w,\lambda) = \frac{1}{2}\|Tu-d\|_Z^2+\frac{1}{2}\alpha\|v\|^2_{L^2(\Omega_v)}+(w,Au+Bv)_U-(\lambda,v)_{L^2(\Omega_v)}. \quad (6)$$

The standard optimality theory states that if (v^*,u^*) is a solution of (1)-(3), then there exist duality functions (w^*,λ^*) such that the Fréchet derivatives of (6), with respect to v, u and w,

$$\left\langle \frac{\partial \mathcal{L}}{\partial v}, \phi \right\rangle = (\alpha v,\phi)_{L^2(\Omega_v)} + (B\phi,w)_U - (\lambda,\phi)_{L^2(\Omega_v)},\ \forall \phi \in L^2(\Omega_v),$$
$$\left\langle \frac{\partial \mathcal{L}}{\partial u}, \phi \right\rangle = (Tu-d,T\phi)_Z + (A\phi,w)_U,\ \forall \phi \in U,$$
$$\left\langle \frac{\partial \mathcal{L}}{\partial w}, \phi \right\rangle = (Au+Bv,\phi)_U,\ \forall \phi \in U,$$

should all be equal to zero at the optimal point (v^*,u^*,w^*,λ^*). In addition, the conditions given by

$$(\lambda v)(x) = 0, \quad (7)$$
$$\lambda(x), v(x) \geq 0, \quad (8)$$

should also be satisfied at this optimal point. By writing the Fréchet derivatives on block form, we get the well known KKT system

$$\begin{bmatrix} \alpha I & 0 & B^* \\ 0 & T^*T & A^* \\ B & A & 0 \end{bmatrix} \begin{bmatrix} v \\ u \\ w \end{bmatrix} = \begin{bmatrix} \lambda \\ T^*d \\ 0 \end{bmatrix}, \quad (9)$$

U onto its dual space U', and hence $\mathcal{A}\mathbf{1}$ is not fulfilled. This can, nevertheless, easily be rectified by applying the inverse Riesz map $R_U^{-1} : U' \to U$ to (2) and thereby obtain the operator $R_U^{-1}A : U \to U$. In this context, one might consider R_U^{-1} to be a preconditioner. We will return to this issue in the example sections.

which we combine with (7)-(8) to obtain the full optimality system. Note that, since we have a convex problem, a solution $(v^*, u^*, w^*, \lambda^*)$ of (7)-(9) will also be a solution of (1)-(3).

4 Primal-dual active set method

To solve our optimization problem, we will follow the primal-dual technique introduced in [2], and later used in [4] and [14].

Thus, we start by noting that (7)-(8) are equivalent to the condition

$$\lambda + \min(0, cv - \lambda) = 0 \quad \forall c > 0.$$

This motivates the PDAS algorithm, where we can define the active $\mathcal{A}$ and inactive $\mathcal{I}$ sets as follows

$$\mathcal{A} = \{x \in \Omega_v : (cv - \lambda)(x) < 0\}, \tag{10}$$

$$\mathcal{I} = \Omega_v \setminus \mathcal{A}, \tag{11}$$

where Ω_v is the domain of the control v. We can now formulate the PDAS method for solving our optimality problem (1)-(3). In the iterative procedure, we need to solve systems on the form (9) at each step, i.e., solve

$$\begin{bmatrix} \alpha I & 0 & B^* \\ 0 & T^*T & A^* \\ B & A & 0 \end{bmatrix} \begin{bmatrix} v^k \\ u^k \\ w^k \end{bmatrix} = \begin{bmatrix} \lambda^k \\ T^*d \\ 0 \end{bmatrix}, \tag{12}$$

together with

$$\lambda^k(x) = 0 \ \text{ on } \mathcal{I}^k, \tag{13}$$

$$v^k(x) = 0 \ \text{ on } \mathcal{A}^k. \tag{14}$$

Note that the unknowns are v^k, u^k, w^k and λ^k, and hence there are unknowns on both sides of equation (12). Here, $\mathcal{A}^k$ and $\mathcal{I}^k$ are the active and inactive sets associated with the kth iteration of the PDAS algorithm, see steps 9 and 10 in Algorithm 1.

In [2] it is shown that the primal-dual active set method provides a local minimum if the active set stays unchanged in two consecutive iterations. We can now, schematically, present the PDAS algorithm, see Algorithm 1.

Although the algorithm is in place, it is possible to reduce the CPU cost of solving (12) - (14). The idea is based on the fact that, at each iteration, we know that the control parameter v^k is zero on the active domain (14), and similarly, we know that the Lagrange multiplier λ^k is zero on the inactive domain (13). Hence, it intuitively seems possible to restrict the control v^k to the inactive domain. Similarly, we want to restrict the Lagrange multiplier λ^k to the active domain. By restricting these functions, the optimality system to be solved becomes smaller, in the sense of fewer indices in the corresponding discretized KKT equations, and hence it will be faster to solve.

Algorithm 1 Primal-dual active-set method

Choose the initial set $\mathcal{A}^0$ of active constraints
$\mathcal{I}^0 = \Omega_v \setminus \mathcal{A}^0$
for k = 0, 1, 2,... **do**
  if $k > 0$ and $\mathcal{A}^k = \mathcal{A}^{k-1}$ **then**
    STOP (algorithm converged)
  else
    Solve (12) - (14)
  end if
  $\mathcal{A}^{k+1} = \{x \in \Omega_v : (cv^k - \lambda^k)(x) < 0\}$
  $\mathcal{I}^{k+1} = \Omega_v \setminus \mathcal{A}^{k+1}$
end for

5 Reduced KKT system

We will now first derive a linear system which only involves the restrictions of v^k and λ^k to the inactive and active domains, respectively. Thereafter, we analyze whether assumptions $\mathcal{A}\mathbf{1}$-$\mathcal{A}\mathbf{5}$, see Section 2, are inherited by this system.

Let $q \in L^2(\Omega_v)$ be arbitrary. We may split $q \in L^2(\Omega_v)$,

$$q(x) = \begin{cases} q^{\mathcal{I}^k}(x) & \text{if } x \in \mathcal{I}^k, \\ q^{\mathcal{A}^k}(x) & \text{if } x \in \mathcal{A}^k. \end{cases} \tag{15}$$

where

$$q^{\mathcal{I}^k} = q|_{\mathcal{I}^k},$$
$$q^{\mathcal{A}^k} = q|_{\mathcal{A}^k}.$$

Let us also introduce the notation

$$L^2(\mathcal{I}^k) = \{q|_{\mathcal{I}^k} : q \in L^2(\Omega_v)\}, \tag{16}$$
$$L^2(\mathcal{A}^k) = \{q|_{\mathcal{A}^k} : q \in L^2(\Omega_v)\},$$

and note that

$$q^{\mathcal{I}^k} \in L^2(\mathcal{I}^k),$$
$$q^{\mathcal{A}^k} \in L^2(\mathcal{A}^k).$$

To derive the reduced KKT system, we need an operator which maps the restricted function $v^{\mathcal{I}^k}$ of the control v^k into the entire control space $L^2(\Omega_v)$. This operator must map a function defined on the domain $\mathcal{I}^k$ into a function defined on the domain Ω_v by employing a zero extension. We will denote this operator by

$$E^{\mathcal{I}^k} : L^2(\mathcal{I}^k) \to L^2(\Omega_v). \tag{17}$$

Note that, for any $r \in L^2(\mathcal{I}^k)$,

$$\left(E^{\mathcal{I}^k} r\right)(x) = r(x) \quad \text{for all } x \in \mathcal{I}^k, \tag{18}$$

$$\left(E^{\mathcal{I}^k} r\right)(x) = 0 \quad \text{for all } x \in \mathcal{A}^k. \tag{19}$$

We also need a similar operator $E^{\mathcal{A}^k}$ for the Lagrange multiplier λ^k. That is, an operator which maps the restricted version $\lambda^{\mathcal{A}^k}$ of λ^k into the full domain Ω_v, by a zero extension. Formally, this is defined as

$$E^{\mathcal{A}^k} : L^2(\mathcal{A}^k) \to L^2(\Omega_v),$$

where this mapping satisfies

$$\left(E^{\mathcal{A}^k} r\right)(x) = r(x) \quad \text{for all } x \in \mathcal{A}^k, \tag{20}$$

$$\left(E^{\mathcal{A}^k} r\right)(x) = 0 \quad \text{for all } x \in \mathcal{I}^k, \tag{21}$$

which holds for any $r \in L^2(\mathcal{A}^k)$. From (18)-(19) and (20)-(21), we can define the inner products of the "restricted" spaces $L^2(\mathcal{I}^k)$ and $L^2(\mathcal{A}^k)$ as

$$(q, r)_{L^2(\mathcal{I}^k)} = (E^{\mathcal{I}^k} q, E^{\mathcal{I}^k} r)_{L^2(\Omega_v)}, \tag{22}$$

$$(q, r)_{L^2(\mathcal{A}^k)} = (E^{\mathcal{A}^k} q, E^{\mathcal{A}^k} r)_{L^2(\Omega_v)}. \tag{23}$$

By construction, $\mathcal{I}^k \cap \mathcal{A}^k = \emptyset$, and (19) and (21) therefore imply that the ranges of $E^{\mathcal{I}^k}$ and $E^{\mathcal{A}^k}$ are orthogonal sets in $L^2(\Omega_v)$,

$$R\left(E^{\mathcal{I}^k}\right) \perp R\left(E^{\mathcal{A}^k}\right). \tag{24}$$

Also note that $E^{\mathcal{I}^k}$ and $E^{\mathcal{A}^k}$ are one-to-one, but not onto. Due to (18)-(19) and (20)-(21), all $q \in L^2(\Omega_v)$ satisfy

$$q = E^{\mathcal{I}^k} q^{\mathcal{I}^k} + E^{\mathcal{A}^k} q^{\mathcal{A}^k}, \tag{25}$$

cf. the splitting (15).

Recall that the linear operator B maps the control in $L^2(\Omega_v)$ into the state space U, see sections 1 and 2. We can now use (25) to conveniently split this mapping:

$$\begin{aligned} Bq &= BE^{\mathcal{I}^k} q^{\mathcal{I}^k} + BE^{\mathcal{A}^k} q^{\mathcal{A}^k} \\ &= B^{\mathcal{I}^k} q^{\mathcal{I}^k} + B^{\mathcal{A}^k} q^{\mathcal{A}^k}, \end{aligned} \tag{26}$$

where

$$B^{\mathcal{I}^k} = BE^{\mathcal{I}^k} : L^2(\mathcal{I}^k) \to U, \tag{27}$$

$$B^{\mathcal{A}^k} = BE^{\mathcal{A}^k} : L^2(\mathcal{A}^k) \to U, \tag{28}$$

With these operators at hand, we are now able to simplify the optimality system (12) - (14). We start with formulating the following lemma.

Lemma 5.1. *Let $E^{\mathcal{I}^k}$ and $E^{\mathcal{A}^k}$ be the extension operators introduced in (18)-(19) and (20)-(21), respectively. Then*

(i) $q = E^{\mathcal{I}^k} q^{\mathcal{I}^k} + E^{\mathcal{A}^k} q^{\mathcal{A}^k}$ *for any* $q \in L^2(\Omega_v)$,

(ii) $Bq = B^{\mathcal{I}^k} q^{\mathcal{I}^k} + B^{\mathcal{A}^k} q^{\mathcal{A}^k}$ *for any* $q \in L^2(\Omega_v)$,

(iii) $B^* = E^{\mathcal{I}^k}[B^{\mathcal{I}^k}]^* + E^{\mathcal{A}^k}[B^{\mathcal{A}^k}]^*$,

where $B^{\mathcal{I}^k}$ and $B^{\mathcal{A}^k}$ are defined in (27) and (28), respectively.

Proof. (i) was established in the derivation leading to (25).

(ii) was established in the derivation leading to (26).

(iii) can be verified has follows. First, (18)-(19) and (20)-(21) imply that, for any $q, r \in L^2(\Omega_v)$,

$$\begin{aligned}(q^{\mathcal{I}^k}, r^{\mathcal{I}^k})_{L^2(\mathcal{I}^k)} &= (q, E^{\mathcal{I}^k} r^{\mathcal{I}^k})_{L^2(\Omega_v)},\\ (q^{\mathcal{A}^k}, r^{\mathcal{A}^k})_{L^2(\mathcal{A}^k)} &= (q, E^{\mathcal{A}^k} r^{\mathcal{A}^k})_{L^2(\Omega_v)}.\end{aligned}$$

Consequently, for arbitrary $q \in L^2(\Omega_v)$ and $s \in U$,

$$\begin{aligned}(q, B^* s)_{L^2(\Omega_v)} &= (Bq, s)_U\\ &= (B^{\mathcal{I}^k} q^{\mathcal{I}^k} + B^{\mathcal{A}^k} q^{\mathcal{A}^k}, s)_U\\ &= \left(q^{\mathcal{I}^k}, [B^{\mathcal{I}^k}]^* s\right)_{L^2(\mathcal{I}^k)} + \left(q^{\mathcal{A}^k}, [B^{\mathcal{A}^k}]^* s\right)_{L^2(\mathcal{A}^k)}\\ &= \left(q, E^{\mathcal{I}^k}[B^{\mathcal{I}^k}]^* s\right)_{L^2(\Omega_v)} + \left(q, E^{\mathcal{A}^k}[B^{\mathcal{A}^k}]^* s\right)_{L^2(\Omega_v)}\\ &= \left(q, \left\{E^{\mathcal{I}^k}[B^{\mathcal{I}^k}]^* + E^{\mathcal{A}^k}[B^{\mathcal{A}^k}]^*\right\} s\right)_{L^2(\Omega_v)}.\end{aligned}$$

Hence, it follows that $B^* = E^{\mathcal{I}^k}[B^{\mathcal{I}^k}]^* + E^{\mathcal{A}^k}[B^{\mathcal{A}^k}]^*$, which finishes the proof.

□

Assume that v^k, u^k, w^k and λ^k satisfy (12)-(14), i.e.

$$\alpha v^k + B^* w^k = \lambda^k, \tag{29}$$

$$T^* T u^k + A^* w^k = T^* d, \tag{30}$$

$$B v^k + A u^k = 0, \tag{31}$$

$$\lambda^k = 0 \text{ on } \mathcal{I}^k, \tag{32}$$

$$v^k = 0 \text{ on } \mathcal{A}^k. \tag{33}$$

From properties (i) and (iii) in Lemma 5.1 we find that equation (29) may be written on the form

$$\begin{aligned}\alpha v^k + B^* w^k &= \alpha E^{\mathcal{I}^k} v^{\mathcal{I}^k} + \alpha E^{\mathcal{A}^k} v^{\mathcal{A}^k} + E^{\mathcal{I}^k}[B^{\mathcal{I}^k}]^* w^k + E^{\mathcal{A}^k}[B^{\mathcal{A}^k}]^* w^k \\ &= E^{\mathcal{I}^k}\lambda^{\mathcal{I}^k} + E^{\mathcal{A}^k}\lambda^{\mathcal{A}^k} = \lambda^k.\end{aligned}$$

Since $\lambda^{\mathcal{I}^k} = 0$ and $v^{\mathcal{A}^k} = 0$,

$$\alpha E^{\mathcal{I}^k} v^{\mathcal{I}^k} + E^{\mathcal{I}^k}[B^{\mathcal{I}^k}]^* w^k + E^{\mathcal{A}^k}[B^{\mathcal{A}^k}]^* w^k = E^{\mathcal{A}^k}\lambda^{\mathcal{A}^k}$$

or

$$E^{\mathcal{I}^k}\left\{\alpha v^{\mathcal{I}^k} + [B^{\mathcal{I}^k}]^* w^k\right\} + E^{\mathcal{A}^k}\left\{[B^{\mathcal{A}^k}]^* w^k - \lambda^{\mathcal{A}^k}\right\} = 0. \tag{34}$$

But recall that the ranges of $E^{\mathcal{I}^k}$ and $E^{\mathcal{A}^k}$ are orthogonal, cf. (24), and that these operators are one-to-one. Consequently, we find that (34) can be split into two equations

$$\begin{aligned}\alpha v^{\mathcal{I}^k} + [B^{\mathcal{I}^k}]^* w^k &= 0, \\ [B^{\mathcal{A}^k}]^* w^k - \lambda^{\mathcal{A}^k} &= 0,\end{aligned}$$

which implies that (29) can be replaced with these two expressions.

Next, we can use property (ii) in Lemma 5.1 to express equation (31) as

$$Bv^k + Au^k = B^{\mathcal{I}^k} v^{\mathcal{I}^k} + B^{\mathcal{A}^k} v^{\mathcal{A}^k} + Au^k = 0$$

or

$$B^{\mathcal{I}^k} v^{\mathcal{I}^k} + Au^k = 0,$$

where we have used that $v^{\mathcal{A}^k} = 0$.

The KKT system (29)-(33) can therefore be written on the form

$$\begin{aligned}\alpha v^{\mathcal{I}^k} + [B^{\mathcal{I}^k}]^* w^k &= 0, \\ [B^{\mathcal{A}^k}]^* w^k - \lambda^{\mathcal{A}^k} &= 0, \\ T^*Tu^k + A^* w^k &= T^* d, \\ B^{\mathcal{I}^k} v^{\mathcal{I}^k} + Au^k &= 0,\end{aligned}$$

Proposition 5.2. *Assume that v^k, u^k, w^k and λ^k solve (12)-(14). Then $v^{\mathcal{I}^k} = v^k|_{\mathcal{I}^k}$, u^k, w^k and $\lambda^{\mathcal{A}^k} = \lambda^k|_{\mathcal{A}^k}$ satisfy*

$$\underbrace{\begin{bmatrix} \alpha I^{\mathcal{I}^k} & 0 & [B^{\mathcal{I}^k}]^* \\ 0 & T^*T & A^* \\ B^{\mathcal{I}^k} & A & 0 \end{bmatrix}}_{=\mathcal{B}_\alpha^k} \begin{bmatrix} v^{\mathcal{I}^k} \\ u^k \\ w^k \end{bmatrix} = \begin{bmatrix} 0 \\ T^*d \\ 0 \end{bmatrix}, \tag{35}$$

$$\lambda^{\mathcal{A}^k} = [B^{\mathcal{A}^k}]^* w^k. \tag{36}$$

With other words, in each iteration of the PDAS method we can solve the block system (35) and thereafter use the straightforward update (36) for the Lagrange multiplier.

6 Spectrum of the reduced KKT system

Assume that assumptions $\mathcal{A}\mathbf{1}$-$\mathcal{A}\mathbf{5}$ hold, see Section 2. In the introduction we mentioned that for a sound discretization of (4), associated with (1)-(2), without the inequality constraint (3), the discrete operator $\mathcal{B}_\alpha^h$ has a spectrum of the form (5). This issue is analyzed in detail in [11]. Krylov subspace solvers therefore handle (4) very well. We have shown in the derivation leading to (35) that we get KKT systems very similar to (4) in each iteration of the PDAS algorithm. One might therefore hope that the MINRES method also is a fast solver for the reduced system (35). This issue can be investigated by exploring whether the operators appearing in $\mathcal{B}_\alpha^k$, defined in (35), also satisfy assumptions $\mathcal{A}\mathbf{1}$-$\mathcal{A}\mathbf{5}$. In short, are these properties, assumed to hold for $\mathcal{B}_\alpha$, inherited by $\mathcal{B}_\alpha^k$? If this is the case, then the spectrum of $\mathcal{B}_\alpha^k$ also will consist of three bounded intervals with a few isolated eigenvalues, i.e. be of the form (5), and Krylov solvers will handle (35) well.

We start by pointing out that (35) is the KKT system associated with the following optimization problem:

$$\min_{(v^{\mathcal{I}^k},u)\in L^2(\mathcal{I}^k)\times U} \left\{ \frac{1}{2}\|Tu-d\|_Z^2 + \frac{1}{2}\alpha\|v^{\mathcal{I}^k}\|_{L^2(\mathcal{I}^k)}^2 \right\}, \tag{37}$$

subject to

$$Au = -B^{\mathcal{I}^k} v^{\mathcal{I}^k} = -BE^{\mathcal{I}^k} v^{\mathcal{I}^k}, \tag{38}$$

where $L^2(\mathcal{I}^k)$, $E^{\mathcal{I}^k}$ and $B^{\mathcal{I}^k}$ are defined in the previous section.

We note that (37)-(38) is on the same form as (1)-(2), except that B in (2) has been replaced with $B^{\mathcal{I}^k} = BE^{\mathcal{I}^k}$. Since the operators A and T are unchanged in the reduced problem (37)-(38), we immediately conclude that (35) fulfills assumptions $\mathcal{A}\mathbf{1}$, $\mathcal{A}\mathbf{2}$, and $\mathcal{A}\mathbf{4}$. It remains to explore $\mathcal{A}\mathbf{3}$ and $\mathcal{A}\mathbf{5}$.

Note that assumption $\mathcal{A}\mathbf{3}$ no longer concerns the operator B, but instead the operator

$$B^{\mathcal{I}^k} = BE^{\mathcal{I}^k} : L^2(\mathcal{I}^k) \to U,$$

cf. the derivation leading to (27). Thus, we must prove that

$$E^{\mathcal{I}^k} : L^2(\mathcal{I}^k) \to L^2(\Omega_v),$$

see (17)-(19), is a bounded and linear operator. It is obvious that such an extension operator is linear, and from (18)-(19) and (22) we find that

$$\|E^{\mathcal{I}^k} r\|_{L^2(\Omega_v)} = \|r\|_{L^2(\mathcal{I}^k)} \quad \text{for any } r \in L^2(\mathcal{I}^k),$$

and therefore

$$\|E^{\mathcal{I}^k}\| = \sup_{r\in L^2(\mathcal{I}^k)} \frac{\|E^{\mathcal{I}^k} r\|_{L^2(\Omega_v)}}{\|r\|_{L^2(\mathcal{I}^k)}} = 1. \tag{39}$$

Since B is assumed to be bounded and linear, we can conclude that $B^{\mathcal{I}^k}$ is linear and bounded, i.e. (35) satisfies assumption $\mathcal{A}\mathbf{3}$.

Although we assumed that (1)-(3) is ill-posed without regularization $\alpha = 0$, see assumption $\mathcal{A}\mathbf{5}$ in Section 2, this may not be the case for (37)-(38) (with $\alpha = 0$). For example, if the inactive set $\mathcal{I}^k$ only contains one element/index, then (37)-(38) typically will be well-posed even with zero regularization. Hence, one can in general not assure that $\mathcal{A}\mathbf{5}$, assumed to be satisfied by $\mathcal{B}_0$, is inherited by $\mathcal{B}_0^k$. There are two possibilities:

- If, luckily, (37)-(38) is well posed for $\alpha = 0$, then regularization is not needed, and the effective numerical solution of this linear system with the MINRES method follows from standard theory.

- If $\mathcal{A}\mathbf{5}$ is inherited by (37)-(38), then $\mathcal{A}\mathbf{1}$-$\mathcal{A}\mathbf{5}$ are satisfied, and a sound discretization $\mathcal{B}_\alpha^{k,h}$ of $\mathcal{B}_\alpha^k$ will have eigenvalues satisfying

$$\text{sp}(\mathcal{B}_\alpha^{k,h}) \subset [-b,-a] \cup [c\alpha, 2\alpha] \cup \{\lambda_1, \lambda_2, ..., \lambda_{N(\alpha)}\} \cup [a,b]. \tag{40}$$

  (Of course, the constants in this expression may differ from those in (5)). From this result, and the Chebyschev polynomial analysis presented in [11], it follows that the number of MINRES iterations needed to solve (35) can not grow faster than of order $O([\ln(\alpha^{-1})]^2)$ as $\alpha \to 0$. Moreover, in practical computations one often observes iterations counts of order $O(\ln(\alpha^{-1}))$. (The latter issue is also discussed from a theoretical point of view in [11]).

Definition 6.1 ("Sound discretization")**.** *A "sound discretization" of $\mathcal{B}_\alpha^k$ means that also the discrete problem should satisfy $\mathcal{A}\mathbf{1} - \mathcal{A}\mathbf{4}$, with operator norms which are bounded independently of the mesh parameter h. In addition, a discrete version of $\mathcal{A}\mathbf{5}$ should hold, i.e. that the eigenvalues of $\mathcal{B}_0^{k,h}$ satisfy*

$$|\lambda_i(\mathcal{B}_0^{k,h})| \leq \tilde{c}e^{-\tilde{C}i}, \quad i = 1, ..., n, \tag{41}$$

where $\tilde{c}, \tilde{C}$ are positive constants.

Remark 6.2. *For finite dimensional problems, there obviously always exist $\tilde{c}$ and $\tilde{C}$ such that* (41) *holds. Our results are therefore only of relevance for problems where*

$$\tilde{c}e^{-\tilde{C}n}$$

is extremely close to zero. That is, much smaller than typical choices of the size of the regularization parameter α. The latter will typically be the case if an ill-posed problem is discretized.

Theorem 6.3. *Let $\mathcal{B}_\alpha^k$ be the operator defined in (35). Assume that assumption $\mathcal{A}\mathbf{5}$ is inherited by* (37)-(38)*. Then, for every $\alpha > 0$ and for a*

sound discretization $\mathcal{B}_\alpha^{k,h}$ *of* $\mathcal{B}_\alpha^k$*, in the sense of Definition 6.1, the spectrum of the associated discretized operator obeys*

$$\mathrm{sp}(\mathcal{B}_\alpha^{k,h}) \subset [-b,-a] \cup [c\alpha, 2\alpha] \cup \{\lambda_1, \lambda_2, ..., \lambda_{N(\alpha)}\} \cup [a,b].$$

Here, a*,* b*, and* c *are positive constants independent of* α *and* $N(\alpha) = O(\ln(\alpha^{-1}))$*.*

Since the operators appearing in $\mathcal{B}_\alpha^k$ fulfill assumptions $\mathcal{A}\mathbf{1}$-$\mathcal{A}\mathbf{5}$, the proof of this theorem is identical to the analysis presented in [11], and therefore omitted.

We conclude, at least theoretically, that the MINRES algorithm is well suited for solving the KKT system (35) appearing in each iteration of the PDAS algorithm applied to the box constrained optimization problem (1)-(3). We will illuminate these findings below with numerical experiments.

7 Example 1

In our first model problem we define

$$\begin{aligned}\Omega &= (0,1) \times (0,1),\\ \Omega_v &= \left(\frac{1}{4}, \frac{3}{4}\right) \times \left(\frac{1}{4}, \frac{3}{4}\right),\end{aligned}$$

and consider the minimization problem

$$\min_{(v,u)\in L^2(\Omega_v)\times H^1(\Omega)} \left\{ \frac{1}{2}\|Tu - d\|^2_{L^2(\partial\Omega)} + \frac{1}{2}\alpha\|v\|^2_{L^2(\Omega_v)} \right\} \tag{42}$$

subject to

$$-\Delta u + u = \begin{cases} -v & \text{if } x \in \Omega_v, \\ 0 & \text{if } x \in \Omega \setminus \Omega_v, \end{cases} \tag{43}$$

$$\nabla u \cdot n = 0 \text{ on } \partial\Omega, \tag{44}$$

$$v(x) \geq 0 \text{ a.e.} \tag{45}$$

Here, T denotes the trace operator $T : H^1(\Omega) \to L^2(\partial\Omega)$, which is well known to bounded and linear, i.e. assumption $\mathcal{A}\mathbf{4}$ holds. Note that the state space U and the observation space Z are

$$U = H^1(\Omega), \tag{46}$$

$$Z = L^2(\partial\Omega). \tag{47}$$

We are thus trying to recover the function $v \in L^2(\Omega_v)$ from an observation of u along the boundary $\partial\Omega$ of Ω.

Remark 7.1. *We want to derive the optimality system associated with (42)-(45) and to solve it with Algorithm 1. There are, however, two issues that must be handled before we can employ the theoretical considerations presented above:*

(a) In the generic state equation (2) we assumed that the operator A is a mapping from the state space U onto the state space U, i.e. $A : U \to U$. This differs from standard PDE theory. For example, the weak form of (43) involves an operator $\widehat{A}$ mapping $H^1(\Omega)$ onto its dual space $(H^1(\Omega))'$.

(b) In order to solve the KKT system associated with (42)-(45) numerically, we must discretize the operators by applying, e.g., the Finite Element Method (FEM).

Both of these matters can be handled adequately, and we will discuss each of them in some detail. It is, however, difficult to treat both problems simultaneously. Therefore, we address them separately, starting with (a), which will provide us with a suitable preconditioner for the continuous KKT system. Thereafter, we briefly comment the discretization of the preconditioned optimality system, i.e. issue (b).

7.1 Preconditioner

Let us explore issue (a). As mentioned above, the discussion of this matter will provide us with a suitable preconditioner for the KKT system arising in each iteration of the PDAS algorithm applied to solve (42)-(45).

The variational form of (43)-(44) reads: Find $u \in U = H^1(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla \psi + u\psi \, dx = -\int_{\Omega_v} v\psi \, dx \quad \text{for all } \psi \in U,$$

or

$$\langle \widehat{A}u, \psi \rangle = -\langle \widehat{B}v, \psi \rangle \quad \text{for all } \psi \in U, \tag{48}$$

where

$$\widehat{A} : U \to U', \quad u \to \int_\Omega \nabla u \cdot \nabla \psi + u\psi \, dx, \; \psi \in U,$$
$$\widehat{B} : L^2(\Omega_v) \to U', \quad v \to \int_{\Omega_v} v\psi \, dx, \; \psi \in U.$$

We may write (48) more compactly, i.e.

$$\widehat{A}u = -\widehat{B}v.$$

In order to obtain an equation of the form (2), where $A : U \to U$ and $B : L^2(\Omega_v) \to U$, we can simply invoke the inverse R_U^{-1} of the Riesz map $R_U : U \to U'$, i.e.

$$R_U^{-1}\widehat{A}u = -R_U^{-1}\widehat{B}v,$$

which is on the desired form since

$$A = R_U^{-1}\widehat{A} : U \to U, \tag{49}$$
$$B = R_U^{-1}\widehat{B} : L^2(\Omega_v) \to U. \tag{50}$$

From standard theory for elliptic PDEs, it follows that A, A^{-1} and B are bounded. We thus conclude that assumptions $\mathcal{A}\mathbf{1}$, $\mathcal{A}\mathbf{2}$ and $\mathcal{A}\mathbf{3}$ are satisfied.

Recall that, in each iteration of the PDAS method, we must solve the system (35). We will now explore the form of this system for the present model problem. In (35),

$$B^{\mathcal{I}^k} = BE^{\mathcal{I}^k},$$

see the discussion leading to (27). In the present context, we may use (50) to write this operator on the form

$$\begin{aligned} B^{\mathcal{I}^k} &= R_U^{-1}\widehat{B}E^{\mathcal{I}^k} \\ &= R_U^{-1}\widehat{B}^{\mathcal{I}^k}, \end{aligned} \tag{51}$$

where we define

$$\widehat{B}^{\mathcal{I}^k} = \widehat{B}E^{\mathcal{I}^k}.$$

Equation (35) also involves the adjoint operators A^* and $[B^{\mathcal{I}^k}]^*$ of A and $B^{\mathcal{I}^k}$. According to a rather technical argument presented in [11],

$$A^* = R_U^{-1}\widehat{A}', \tag{52}$$
$$[B^{\mathcal{I}^k}]^* = [R_{L^2(\mathcal{I}^k)}]^{-1}[\widehat{B}^{\mathcal{I}^k}]', \tag{53}$$

where the "$'$" notation is used to denote dual operators, and $R_{L^2(\mathcal{I}^k)}$ is the Riesz map of $L^2(\mathcal{I}^k)$ to its dual space, see (16).

From (49), (51), (52) and (53) it follows that the operator $\mathcal{B}_\alpha^k$ in (35) can be written on the form

$$\begin{aligned} \mathcal{B}_\alpha^k &= \begin{bmatrix} \alpha I^{\mathcal{I}^k} & 0 & [B^{\mathcal{I}^k}]^* \\ 0 & T^*T & A^* \\ B^{\mathcal{I}^k} & A & 0 \end{bmatrix} \\ &= \begin{bmatrix} \alpha I^{\mathcal{I}^k} & 0 & [R_{L^2(\mathcal{I}^k)}]^{-1}[\widehat{B}^{\mathcal{I}^k}]' \\ 0 & T^*T & R_U^{-1}\widehat{A}' \\ R_U^{-1}\widehat{B}^{\mathcal{I}^k} & R_U^{-1}\widehat{A} & 0 \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} [R_{L^2(\mathcal{I}^k)}]^{-1} & 0 & 0 \\ 0 & R_U^{-1} & 0 \\ 0 & 0 & R_U^{-1} \end{bmatrix}}_{=[\mathcal{R}^k]^{-1}} \underbrace{\begin{bmatrix} \alpha R_{L^2(\mathcal{I}^k)} & 0 & [\widehat{B}^{\mathcal{I}^k}]' \\ 0 & R_U T^*T & \widehat{A}' \\ \widehat{B}^{\mathcal{I}^k} & \widehat{A} & 0 \end{bmatrix}}_{=\widehat{\mathcal{B}}_\alpha^k}. \end{aligned} \tag{54}$$

We can therefore express

$$\mathcal{B}_\alpha^k p^k = b,$$

cf. (35), appearing in each iteration of the PDAS algorithm, as

$$\begin{bmatrix} [R_{L^2(\mathcal{I}^k)}]^{-1} & 0 & 0 \\ 0 & R_U^{-1} & 0 \\ 0 & 0 & R_U^{-1} \end{bmatrix} \begin{bmatrix} \alpha R_{L^2(\mathcal{I}^k)} & 0 & [\widehat{B}^{\mathcal{I}^k}]' \\ 0 & R_U T^* T & \widehat{A}' \\ \widehat{B}^{\mathcal{I}^k} & \widehat{A} & 0 \end{bmatrix} \begin{bmatrix} v^{\mathcal{I}^k} \\ u^k \\ w^k \end{bmatrix}$$
$$= \begin{bmatrix} [R_{L^2(\mathcal{I}^k)}]^{-1} & 0 & 0 \\ 0 & R_U^{-1} & 0 \\ 0 & 0 & R_U^{-1} \end{bmatrix} \begin{bmatrix} 0 \\ R_U T^* d \\ 0 \end{bmatrix}. \qquad (55)$$

Written more compactly, this system reads

$$[\mathcal{R}^k]^{-1} \widehat{\mathcal{B}}_\alpha^k p^k = [\mathcal{R}^k]^{-1} \widehat{b}, \qquad (56)$$

where

$$\widehat{b} = \mathcal{R}^k b = \begin{bmatrix} 0 \\ R_U T^* d \\ 0 \end{bmatrix},$$
$$p^k = \begin{bmatrix} v^{\mathcal{I}^k} \\ u^k \\ w^k \end{bmatrix}.$$

Note that

$$\widehat{\mathcal{B}}_\alpha^k : L^2(\mathcal{I}^k) \times U \times U \to \Big(L^2(\mathcal{I}^k) \times U \times U \Big)',$$

and that

$$[\mathcal{R}^k]^{-1} : \Big(L^2(\mathcal{I}^k) \times U \times U \Big)' \to L^2(\mathcal{I}^k) \times U \times U.$$

One may therefore regard $[\mathcal{R}^k]^{-1}$ to be a preconditioner for the (continuous) KKT system arising in each iteration of the PDAS method applied to (42)-(45), see [9] for further details. Note that the operators $\mathcal{R}^k$, $[\mathcal{R}^k]^{-1}$, $\widehat{\mathcal{B}}_\alpha^k$ and $[\widehat{\mathcal{B}}_\alpha^k]^{-1}$ are bounded. Hence, a proper discretization of these mappings should yield a discretized approximation of (54) which is well behaved for any mesh parameter $h > 0$. This completes the discussion of issue (a).

7.2 Discretization

Let us turn our attention towards the discretization matter mentioned in (b), i.e. the discretization of (56). Recall that $\mathcal{B}_\alpha^k = [\mathcal{R}^k]^{-1} \widehat{\mathcal{B}}_\alpha^k$ only operates on the inactive part of the control. Expressed with mathematical symbols,

$$\mathcal{B}_\alpha^k : L^2(\mathcal{I}^k) \times U \times U \to L^2(\mathcal{I}^k) \times U \times U.$$

Hence, in each iteration of the PDAS method one may regard $L^2(\mathcal{I}^k)$ to be the control space, while the state space U and the observation space Z are defined in (46)-(47), respectively.

As mentioned earlier, one may think of the inverse Riesz maps $[R_{L^2(\mathcal{I}^k)}]^{-1}$ and R_U^{-1}, see (54), as preconditioners. Since $U = H^1(\Omega)$, it follows that, in a FEM setting,

- $R_{L^2(\mathcal{I}^k)}$ "corresponds" to the mass matrix $M_v^{\mathcal{I}^k,\mathcal{I}^k}$ associated with the inactive set $\mathcal{I}^k \subset \Omega_v$,
- R_U "corresponds" to the sum of the mass matrix M and the stiffness matrix S associated with the domain Ω.

Concerning the details of the discretization of the operators in $\widehat{\mathcal{B}}_\alpha^k$, defined in (54), we refer to [9]. If we use the superscript notation "$\mathcal{I}^k$" and " : " to denote the inactive indices and all the indices, respectively, the end result is as follows:

- $\widehat{A}$ yields the matrix $M + S$, which is the sum of the mass and stiffness matrix associated with the domain Ω.
- $\widehat{B}^{\mathcal{I}^k}$ yields the matrix $M_v^{\mathcal{I}^k,:}$, where M_v is the mass matrix associated with the sub domain Ω_v of Ω.
- $R_U T^* T$ yields the matrix M_∂, which is the mass matrix associated with the boundary $\partial\Omega$ of the domain Ω.
- The functions v, u, w and d yields the corresponding vectors $\bar{v}$, $\bar{u}$, $\bar{w}$ and $\bar{d}$.

Hence, the discretized system associated with (55) reads

$$\begin{bmatrix} M_v^{\mathcal{I}^k,\mathcal{I}^k} & 0 & 0 \\ 0 & M+S & 0 \\ 0 & 0 & M+S \end{bmatrix}^{-1} \underbrace{\begin{bmatrix} \alpha M_v^{\mathcal{I}^k,\mathcal{I}^k} & 0 & M_v^{\mathcal{I}^k,:} \\ 0 & M_\partial & M+S \\ M_v^{:,\mathcal{I}^k} & M+S & 0 \end{bmatrix}}_{\bar{\mathcal{B}}_\alpha^k} \underbrace{\begin{bmatrix} \bar{v}^{\mathcal{I}^k} \\ \bar{u}^k \\ \bar{w}^k \end{bmatrix}}_{\bar{p}^k}$$
$$= \begin{bmatrix} M_v^{\mathcal{I}^k,\mathcal{I}^k} & 0 & 0 \\ 0 & M+S & 0 \\ 0 & 0 & M+S \end{bmatrix}^{-1} \underbrace{\begin{bmatrix} 0 \\ M_\partial \bar{d} \\ 0 \end{bmatrix}}_{\bar{b}}. \tag{57}$$

We thus use the preconditioner

$$[\bar{\mathcal{R}}^k]^{-1} = \begin{bmatrix} M_v^{\mathcal{I}^k,\mathcal{I}^k} & 0 & 0 \\ 0 & M+S & 0 \\ 0 & 0 & M+S \end{bmatrix}^{-1}. \tag{58}$$

We have now handled both issues (a) and (b), and derived a discretized preconditioned KKT system (57). It remains to discretize the Lagrange multiplier update (36). Since the procedure for doing this is very similar to the discussion of the KKT system, we leave the technical details to Appendix A. The end result is the update

$$M_v^{\mathcal{A}^k,\mathcal{A}^k}\bar{\lambda}^{\mathcal{A}^k} = M^{\mathcal{A}^k,:}\bar{w}^k, \tag{59}$$

where " $\mathcal{A}^k$ " denotes the active indices.

To summarize, in each iteration of the PDAS algorithm we must solve the preconditioned system (57). The Lagrange multiplier $\bar{\lambda}^{\mathcal{A}^k}$ is thereafter computed by solving (59). Finally, the active and inactive sets are updated according to steps 9 and 10 in Algorithm 1.

7.3 Numerical setup

- All code was written in the framework of cbc.block, which is a FEniCS-based Python implemented library for block operators. See [8] for a full description of cbc.block.

- We used the PyTrilinos package to compute an approximation of the preconditioner (58), using algebraic multigrid (AMG) with a symmetric Gauss-Seidel smoother and three smoothing sweeps. All tables containing iteration counts for the MINRES method were generated with this approximate inverse Riesz map. On the other hand, the eigenvalues of the KKT systems $[\bar{\mathcal{R}}^k]^{-1}\bar{\mathcal{B}}_\alpha^k$, see (57)-(58), were computed with an *exact* inverse $[\bar{\mathcal{R}}^k]^{-1}$ computed in Octave.

- We divided the domain of $\Omega = (0,1)\times(0,1)$ into $N \times N$ squares, and each of these squares were divided into two triangles.

- The following stopping criterion was used to stop the MINRES iteration process

$$\frac{\|r_n^k\|}{\|r_0^k\|} = \left[\frac{(\ \bar{\mathcal{B}}_\alpha^k\bar{p}_n^k - \bar{b}, [\bar{\mathcal{R}}^k]^{-1}[\bar{\mathcal{B}}_\alpha^k\bar{p}_n^k - \bar{b}]\ )}{(\ \bar{\mathcal{B}}_\alpha^k\bar{p}_0^k - \bar{b}, [\bar{\mathcal{R}}^k]^{-1}[\bar{\mathcal{B}}_\alpha^k\bar{p}_0^k - \bar{b}]\ )}\right]^{1/2} < \epsilon, \tag{60}$$

  where ϵ is a small positive parameter. Note that the superindex k is the iteration index for the "outer" PDAS method, while the subindex n is the iteration index for the "inner" MINRES algorithm at each step of the PDAS method.

- In the synthetic examples no noise was added to the input data d, see (1). For the problem involving real world data, however, the input data was given by clinical recordings and obviously contained a significantly amount of noise.

- Synthetic observation data d, used in (42), was produced by setting

$$v(x) = 3\sin(2\pi x_1), x = (x_1, x_2) \in \Omega_v, \tag{61}$$

in (43). Thereafter the boundary value problem (43)-(44) was solved and d was put equal to $u|_{\partial\Omega}$. Note that the control (61) cannot be recovered by solving the optimality system (42)-(45), due to the inequality constraint $v(x) \geq 0$. Hence, the problem formulation might seem peculiar, but as the goal of this example is to study the iteration numbers for the reduced KKT systems, it is desirable to have active constraints for all reasonable values of the regularization parameter α. An experimental investigation suggested the use of a control function of the form (61) to obtain nonempty active sets for large values of the regularization parameter α ($\alpha \approx 1$).

7.4 Results

We are now ready to proceed to the actual experiments. In the introduction we mentioned that the KKT system associated with (1)-(2), without box constraints, has a spectrum of the form (5), as long as assumptions $\mathcal{A}\mathbf{1}$-$\mathcal{A}\mathbf{5}$ in Section 2 are fulfilled. Recall that Theorem 6.3 asserts that such a spectrum will be inherited by each subsystem in the PDAS algorithm, provided that assumption $\mathcal{A}\mathbf{5}$ still holds. Figure 1 shows the spectrum of such a subsystem. It is definitely on the form (40), and we expect that the MINRES method will solve the KKT systems efficiently.

Table 1 contains the average number of MINRES iterations required to solve the reduced KKT systems. That is, the average number of MINRES iterations needed in each iteration of the PDAS algorithm. In these experiments we used a zero initial guess in every run of the MINRES method, i.e. $\bar{p}_0^k = 0$, see (60).

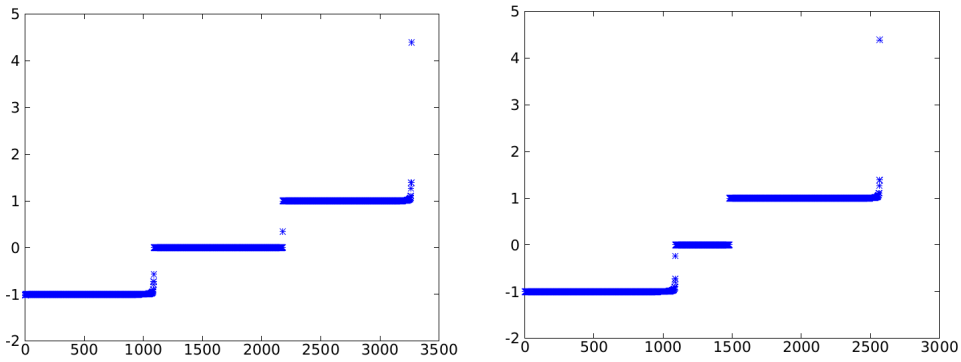
In [11] the authors proved that the number of required MINRES iterations cannot grow faster than $O([\ln(\alpha^{-1})]^2)$, and also explained why iterations counts of order $O([\ln(\alpha^{-1})])$ often will occur in practice. Consider the last row of Table 1, i.e. $N = 512$. For the stopping criterion $\epsilon = 10^{-6}$ in (60), the iteration counts can be relatively well modeled by the formula

$$32.2 - 10.5\log_{10}(\alpha),$$

where we used the method of least squares to estimate the constants in this expression. Similarly, for $N = 512$ and the stopping criterion $\epsilon = 10^{-10}$, we can model the work effort rather accurately with the formula

$$45.0 - 20.1\log_{10}(\alpha).$$

We conclude that the required number of MINRES iteration only grows (approximately) logarithmically as the regularization parameter $\alpha \to 0$.



a) b)

Figure 1: The eigenvalues of $[\bar{\mathcal{R}}^k]^{-1}\bar{\mathcal{B}}_\alpha^k$ in Example 1. Panel a) displays the eigenvalues of the full system, i.e. no active constraints and $\mathcal{I}^k = \Omega_v$. Furthermore, $\alpha = 0.0001$ and $N = 32$. Panel b) shows the spectrum of a reduced KKT system, with 700 active inequalities. We observe that there are fewer eigenvalues in the interval $[c\alpha, 2\alpha]$ in panel b), cf. (40)). More specifically, 700 eigenvalues have been "removed" from this interval in panel b), compared with panel a). We do not present a plot of the isolated eigenvalues, i.e. $\lambda_i \in (2\alpha, a)$, since the full system only has three isolated eigenvalues, and the reduced system only has one isolated eigenvalue.

Note that the spectral condition number $\kappa(\mathcal{B}_\alpha^{k,h})$ of $\mathcal{B}_\alpha^{k,h}$ is of order $O(\alpha^{-1})$, which is "confirmed" by Figure 1. The standard theory for Krylov subspace solvers states that MINRES needs $O(\kappa(\mathcal{B}_\alpha^{k,h}))$ iterations. Hence, the classical estimate provides a very pessimistic estimate for the needed workload.

Table 1 contains iteration counts for both $\epsilon = 10^{-6}$ and $\epsilon = 10^{-10}$, cf. the stopping condition (60). We observe that the iteration numbers increase roughly by a factor of 1.5 if ϵ is decreased from 10^{-6} to 10^{-10}. However, we see no visible difference between the controls v_1 and v_2 computed with these two stopping criteria, see Figure 2. In fact, the relative difference between the solutions depicted in this figure is $2.12 * 10^{-5}$. In retrospect, we conclude that the choice $\epsilon = 10^{-10}$ does not significantly increase the accuracy of the solution compared to the choice $\epsilon = 10^{-6}$. Thus, choosing a suitable stopping criterion is a delicate matter; the criterion must be strict enough to obtain convergence, but not so hard that many unnecessary iterations are performed.

We have previously mentioned that the experiments presented in Table 1 were performed using the zero initial guess in every run of the MINRES method, i.e. $\bar{p}_0^k = 0$. Intuitively, the initial guess $\bar{p}_0^k = \bar{p}_n^{k-1}$ might seem

N\α	1	.1	.01	.001	.0001
32	23	32	38	46	56
64	27	36	42	51	66
128	27	37	42	52	71
256	33	42	48	59	75
512	33	44	52	59	78

(a) Stopping criterion $\epsilon = 10^{-6}$.

N\α	1	.1	.01	.001	.0001
32	34	45	55	70	86
64	39	52	64	83	103
128	41	54	67	85	109
256	48	61	75	95	121
512	49	64	80	103	130

(b) Stopping criterion $\epsilon = 10^{-10}$.

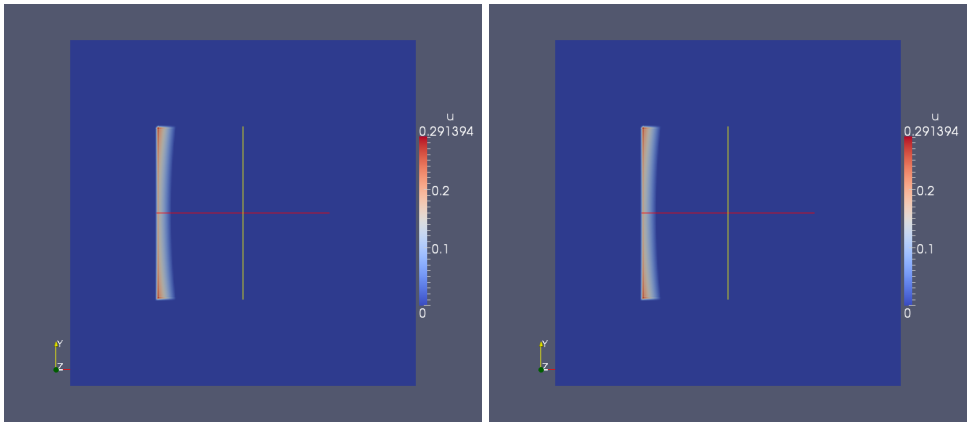
Table 1: The average number of MINRES iterations required to solve the reduced KKT systems in the PDAS algorithm. The two panels display the iteration counts for two different choices of ϵ, see (60). Here, we used the initial guess $\bar{p}_0^k = 0$ in the MINRES algorithm for iteration k of the PDAS method.

preferable. That is, we set the initial guess for the MINRES algorithm equal to the solution from the previous PDAS iteration. In this case, however, (60) should be adjusted to avoid an unreasonable strict stopping criterion when $\bar{p}_n^{k-1} \approx \bar{p}^*$, where $\bar{p}^*$ is the exact solution of the discretized PDE constrained optimization problem. We suggest the following alternative stopping criterion to terminate the MINRES iteration process:

$$\frac{\|r_n^k\|}{\|r_0^0\|} = \left[\frac{(\ \bar{\mathcal{B}}_\alpha^k \bar{p}_n^k - \bar{b}, [\bar{\mathcal{R}}^k]^{-1}[\bar{\mathcal{B}}_\alpha^k \bar{p}_n^k - \bar{b}]\ )}{(\ \bar{\mathcal{B}}_\alpha^0 \bar{p}_0^0 - \bar{b}, [\bar{\mathcal{R}}^0]^{-1}[\bar{\mathcal{B}}_\alpha^0 \bar{p}_0^0 - \bar{b}]\ )} \right]^{1/2} < \epsilon. \tag{62}$$

Note that the initial guess $\bar{p}_0^k = \bar{p}_n^{k-1}$ and the alternative stopping criterion (62) will consistently be used together. Similarly, when we employ the initial guess $\bar{p}_0^k = 0$, the criterion (60) will be used to terminate the iteration process.

How these two different initial guesses affect the iteration counts, can be observed by comparing Table 1 with Table 2. In Table 1 we used the initial guess $\bar{p}_0^k = 0$ in every run of the MINRES method, whereas for the numbers presented in Table 2 we employed $\bar{p}_0^k = \bar{p}_n^{k-1}$. For large values of α, we observe a reduction in the iteration counts, but this effect seems to be less apparent for the smaller values of α. We suspect this to be linked to our choice of synthetic observation data, d, which was generated by the control (61). For this observation data d, and small values of α, the solutions of (42)-(44) and (42)-(45) are very different, i.e. the inequality constraints have a significant impact. As a result of this difference, the initial guess $\bar{p}_0^k = \bar{p}_n^{k-1}$ is not much better than the zero guess. We will return to this matter in the next section.



(a) Stopping criterion $\epsilon = 10^{-6}$. (b) Stopping criterion $\epsilon = 10^{-10}$.

Figure 2: The solution of (42)-(45) for two different stopping criteria. In these examples, $N = 256$ and $\alpha = 0.01$. The relative difference $\frac{\|v_1 - v_2\|_{L^2(\Omega)}}{\|v_1\|_{L^2(\Omega)}}$ between these two control functions is $2.12 * 10^{-5}$.

8 The inverse problem of electrocardiography

We will now study a real world problem. In the *inverse problem of electrocardiography* one attempts to identify an ischemic region/infarction by combining ECG recordings with the, so called, bidomain model [2]. Since the derivation of the bidomain model is not essential for understanding the optimization problem, we refer to [13] for further details about this model.

The control function v in this application, however, must be addressed in some detail. In this medical problem, the control v is the transmembrane potential of the heart, i.e. the potential difference over the cell membrane of the heart cells. According to biomedical knowledge, we know *a priori* that this potential satisfies

$$v(x) \approx \begin{cases} 0mV & x \text{ in healthy tissue,} \\ 50mV & x \text{ in ischemic tissue.} \end{cases} \tag{63}$$

Our objective is to compute the transmembrane potential v by solving an optimization problem. Thereafter, we use (63) to determine the ischemic region, i.e. this region is the sub-domain of the heart where $v(x) \approx 50$.

The optimization problem will be related to the form (1)-(3), where we have the following information:

- The input data d in (1) is a normalized clinical ECG recording.

[2] Ischemia is a state of reduced blood supply to the heart, usually due to coronary artery disease. It is a reversible condition, but also a precursor to a full heart attack.

N\α	1	.1	.01	.001	.0001
32	16	14	36	46	54
64	15	28	36	50	65
128	13	22	31	46	64
256	15	26	35	49	68
512	15	23	36	51	69

(a) Stopping criterion $\epsilon = 10^{-6}$.

N\α	1	.1	.01	.001	.0001
32	27	32	51	70	85
64	25	46	58	80	102
128	27	35	60	80	105
256	32	40	62	79	103
512	25	46	64	90	109

(b) Stopping criterion $\epsilon = 10^{-10}$.

Table 2: The average number of MINRES iterations required to solve the reduced KKT systems in the PDAS algorithm. The two tables contain the iteration counts for two different choices of ϵ, see (62). Here, we used the initial guess $\bar{p}_0^k = \bar{p}_n^{k-1}$ in the MINRES algorithm for iteration k of the PDAS method.

- The state equation (2) will be the bidomain model[3].
- We use (63) to define suitable inequality constraints.
- The *control space*, however, is no longer an L^2-space, but an H^1-space.

In detail, the optimization problem can be formulated as follows

$$\min_{(v,u)\in H^1(\Omega_H)\times H^1(\Omega_B)} \left\{ \frac{1}{2}\|Tu - d\|^2_{L^2(\partial\Omega_B)} + \frac{1}{2}\alpha\|v\|^2_{H^1(\Omega_H)} \right\} \tag{64}$$

subject to

$$\int_{\Omega_B} \nabla\psi \cdot M\nabla u \; dx = -\int_{\Omega_H} \nabla\psi \cdot M_i\nabla v \; dx, \;\; \forall\psi \in X, \tag{65}$$

$$v(x) \geq 0, \;\; x \in \Omega_H, \tag{66}$$

where

$$M(x) \approx \begin{cases} M_i(x) + M_e(x), & x \in \Omega_H, \\ M_o(x), & x \in \Omega_T. \end{cases}$$

Remark 8.1. *Note that (63) also implies an upper bound for v. This upper bound, however, is dependent on a number of model parameters and is, for reasons outside the scope of this article, not as essential as the lower bound. In addition, our simulations did not provoke any active upper constraints.*

In this section we use the following notation:

· v is the transmembrane potential.

[3]As in Example 1, the bidomain equation involves an operator $\hat{A}$ mapping U onto its dual space U'. Hence, we need an inverse Riesz map to obtain a minimization problem of the form (1)-(3).

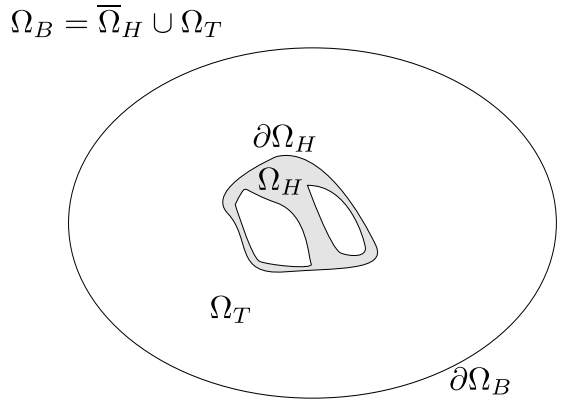


Figure 3: A 2D picture of the domains. Ω_H represents the heart and is depicted in gray color. We denote the remaining domain by the torso, Ω_T. The cavities (white areas) inside the heart represent the ventricles.

- u is the extracellular potential.
- d_{raw} is the ECG recording, and $d = d_{\text{raw}} - \frac{1}{|\partial\Omega_B|}\int_{\partial\Omega_B} d_{\text{raw}}$ is a normalization of the data with respect to the boundary integral, see [10] for details.
- M_i and M_e are the intracellular and extracellular conductivity tensors of the heart, respectively.
- M_o is the extracellular conductivity of the torso.
- Ω_H is the domain of the heart.
- Ω_T is the domain of the torso.
- $\Omega_B = \overline{\Omega}_H \cup \Omega_T$ is the domain of the body.
- $U = \{q \in H^1(\Omega_B) : \int_{\partial\Omega_B} q = 0\}$. Reasons for using this particular Hilbert space are discussed in [10].

For a visual representation of the domains Ω_H, Ω_T and Ω_B, see Figure 3.

Remark 8.2. *In this example, the control space is no longer $L^2(\Omega_v)$, but $H^1(\Omega_H)$, which is not covered by the analysis presented in the theoretical sections. To derive a PDAS algorithm for this H^1-framework is, to the authors knowledge, still an open challenge. Essentially, the problem is that the inequality conditions can no longer be expressed on the simple explicit form* (7)-(8)*, but instead involve solving an obstacle problem, see [6] for further details.*

For a strictly finite dimensional optimization problem, however, a PDAS algorithm exists. Unfortunately, we can then no longer guarantee that it will reflect the structure of the associated infinite dimensional problem. Nevertheless, we find it interesting to investigate the problem from a practical point of view.

Since the discretization of the optimality system associated with (64)-(66) is almost identical to the discretization of the optimality system in Example 1, we will first present the results and thereafter return to the mathematical treatment of (64)-(66).

For the simulations, we have two different sets of patient data, both recorded at Oslo University Hospital. For each of the two patients, we have patient specific geometrical models. Figure 4 shows the body mesh associated with Patient 1. Note that the grid is highly unstructured.

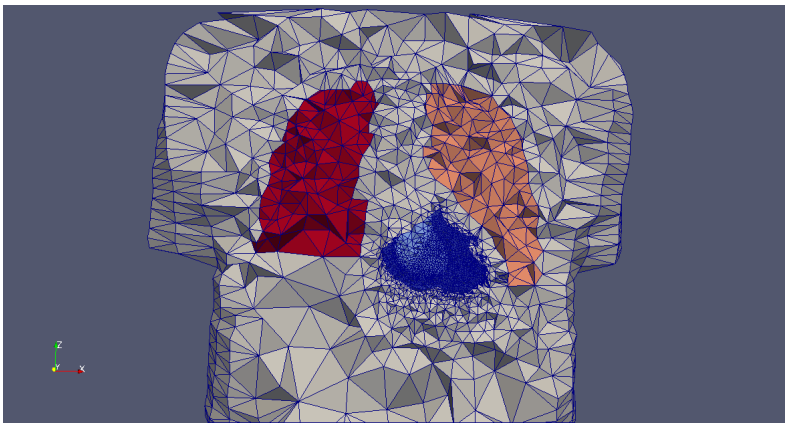


Figure 4: The body mesh associated with Patient 1. The blue color represents the heart, and the red colors represent the lungs. The mesh consists of $51,489$ nodes, whereof 33,156 are located in the heart.

8.1 Results

Table 3 and Table 4 contains the iteration counts for Patient 1 and Patient 2, respectively. The numbers are much higher than those reported for the synthetic example (Example 1), but the growth is still (approximately) logarithmic as $\alpha \to 0$. For Patient 1 the iteration counts for $k = 0$, i.e. the first PDAS iteration, can be modeled by the formula

$$2064.6 - 1287.6 \log_{10}(\alpha).$$

Similarly, we can model the average workload for Patient 1 by the formula

$$1225 - 798.4 \log_{10}(\alpha).$$

We would like to stress that, in this example, the relatively high iteration numbers do not appear to be linked to the fact that our control space is H^1, instead of L^2. More precisely, the iteration counts for $k = 0$, i.e. when there are *no* active constraints, are not lower than for $k > 0$. Other possible explanations for the high iteration numbers will be discussed in Section 9.

For this real world application, we are not only interested in the iteration counts, but also in the actual time it takes to solve the optimization problem. All simulations were performed on a regular laptop with the *Intel®Core^{TM}i5-2520M CPU @ 2.50GHz × 4* processor. From Table 3, we conclude that it lasted between 5 and 13 minutes to solve the inequality constrained optimization problem for Patient 1. For Patient 2, it took between 6 and 15 minutes, depending on the choice of α. For the particular choice of regularization parameter $\alpha = 0.1$, 664 seconds were required. The computed control function for this choice of α can be seen in Figure 5. The figure also displays the solution of (64)-(65), i.e. the optimization problem without the inequality constraint. We see that the introduction of (66) sharpens the image, and thus provides a more well defined separation of the ischemic region and the healthy tissue. For the cardiologists, such a clear distinction is definitely desirable. In fact, one may argue that the image computed without box constraints is of no practical value.

$k\backslash\alpha$	1	$10^{-1/2}$	10^{-1}	$10^{-3/2}$	10^{-2}
0	1808	2851	3694	3911	4497
1	1127	1480	1967	2281	2426
2	361	741	880	1046	1279
Mean	1099	1691	2180	2413	2734
Wall Time	308s	467s	598s	659s	770s

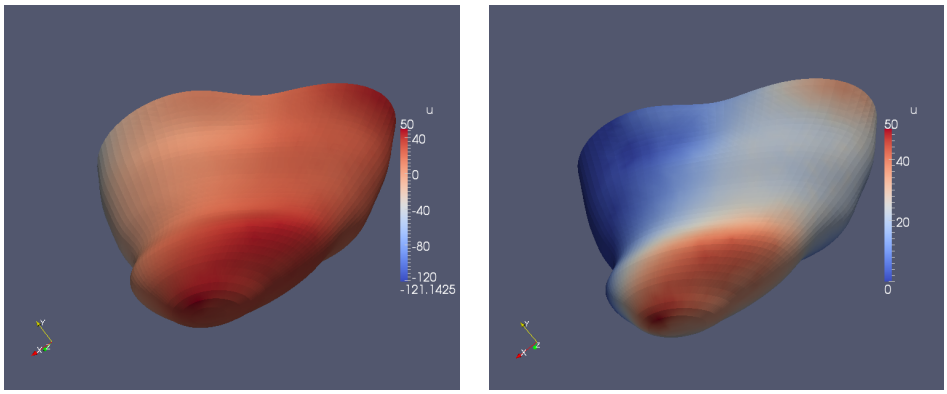
Table 3: The wall time and the number of MINRES iterations required to solve the optimization problem for Patient 1. Note that k denotes the PDAS iteration number. Here, the stopping criterion was $\epsilon = 10^{-6}$, see (62).

$k\backslash\alpha$	1	$10^{-1/2}$	10^{-1}	$10^{-3/2}$	10^{-2}
0	1879	2977	3224	4080	4717
1	1332	1747	2499	2751	3256
2	608	1032	1403	2005	2233
Mean	1273	1919	2375	2945	3402
Wall Time	362s	538s	664s	794s	909s

Table 4: The wall time and the number of MINRES iterations required to solve the optimization problem for Patient 2. Note that k denotes the PDAS iteration number. Here, the stopping criterion was $\epsilon = 10^{-6}$, see (62).

$k\backslash\alpha$	1	$10^{-1/2}$	10^{-1}	$10^{-3/2}$	10^{-2}
Mean	1488	2111	2821	3027	3408
Wall Time	400s	586s	761s	810s	902s

Table 5: The wall time and the average number of MINRES iterations required to solve the optimization problem for Patient 1. These numbers were generated with the initial guess $\bar{p}_0^k = 0$ in every run of the MINRES method, and the stopping criterion was $\epsilon = 10^{-6}$, see (60).



(a) Inverse solution *without* inequality constraints.

(b) Inverse solution *with* inequality constraints.

Figure 5: The computed transmembrane potential v for Patient 2. Here, $\alpha = 0.1$. Panel a) shows the solution of (64)-(65). Panel b), on the other hand, displays the solution of the full problem (64)-(66).

Recall that we, in Example 1, discussed the effect of the initial guess on the performance of the MINRES algorithm. In the present real world application, we have so far reported results obtained with the initial guess $\bar{p}_0^k = \bar{p}_n^{k-1}$. For reason of comparison, we also ran simulations with $\bar{p}_0^k = 0$, see (60). The iteration counts and wall time obtained for these computations can be found in Table 5. Contrary to what was observed in Example 1, we conclude that the initial guess $\bar{p}_0^k = \bar{p}_n^{k-1}$ yields a significant improvement, compared with the "naive" guess $\bar{p}_0^k = 0$. We save roughly $400 - 600$ iterations on average. From a computing-time perspective, the reduction is also significant, with savings in the range of 90 seconds to 3 minutes, i.e. about a 20% reduction in computing-time.

8.2 Discretization

We now return to the mathematical aspects of (64)-(66). Note that the control space V, the state space U and the observation space Z are

$$\begin{aligned} V &= H^1(\Omega_H), \\ U &= \left\{ q \in H^1(\Omega_B) : \int_{\partial\Omega_B} q = 0 \right\}, \\ Z &= L^2(\partial\Omega_B), \end{aligned}$$

see Figure 3 for an overview of the domains. Hence, we are trying to recover a function $v \in H^1(\Omega_H)$ from an observation $d \in L^2(\partial\Omega_B)$ of u along the boundary $\partial\Omega_B$ of the body Ω_B. Notice the form of (63). Since the unknown control is known, a priori, to be approximately piecewise constant, it seems natural to put more weight on the derivative of v in the regularization. Therefore, we use the weighted norm

$$\|v\|_V^2 = \rho\|v\|_{L^2(\Omega_H)}^2 + \|\nabla v\|_{L^2(\Omega_H)}^2$$

on V, where $0 < \rho \ll 1$. This will be reflected in the block operators presented below. In our experiments, we have chosen $\rho = 10^{-4}$.

We start our derivation of the optimality system by considering the state equation (65). This equation can be written as

$$\langle \widehat{A}u, \psi \rangle = -\langle \widehat{B}v, \psi \rangle, \ \forall \psi \in U,$$

where

$$\begin{aligned} \widehat{A} : U \to U', \ \ u \to \int_{\Omega_B} \nabla\psi \cdot M\nabla u \ dx, \psi \in U, \\ \widehat{B} : V \to U', \ \ v \to \int_{\Omega_H} \nabla\psi \cdot M_i\nabla v \ dx, \psi \in U. \end{aligned}$$

We can now proceed as in Example 1 and derive a KKT system with a structure similar to (55). Once more, we refer to [9] for details regarding the matrix representation of the operators in the KKT system. By letting " $\mathcal{I}^k$ " and " : " denote the inactive indices and all indices, respectively, the discretization can roughly be described as follows:

- $R_{V^{\mathcal{I}^k}}$ yields the sum $(\rho M_H + S_H)^{\mathcal{I}^k,\mathcal{I}^k}$ of the mass matrix M_H and the stiffness matrix S_H associated with the domain $\mathcal{I}^k \subset \Omega_H$.
- R_U yields the stiffness matrix S_B associated with the domain Ω_B. [4]

[4] Recall that $U = \{q \in H^1(\Omega_B) : \int_{\partial\Omega_B} q = 0\}$, which makes it possible to use the Poincaré inequality to define the norm $\|\cdot\|_U$ on U as $\|q\|_U = \int_{\Omega_B} |\nabla q|^2$. It therefore follows that the Riesz map only yields the stiffness matrix.

- $\widehat{A}$ yields the matrix N associated with the operator $-\nabla \cdot M \nabla u$ on Ω_B.
- $\widehat{B}$ yields the matrix L associated with the operator $\nabla \cdot M_i \nabla v$ on Ω_H, and consequently, $\widehat{B}^{\mathcal{I}^k}$ yields the matrix $L^{\mathcal{I}^k,:}$.
- $R_U T^* T$ yields the matrix M_∂, which is the mass matrix associated with the boundary $\partial\Omega_B$ of the body Ω_B.

Hence, the discretized KKT system will in this case read:

$$\begin{bmatrix} (\rho M_H + S_H)^{\mathcal{I}^k,\mathcal{I}^k} & 0 & 0 \\ 0 & S_B & 0 \\ 0 & 0 & S_B \end{bmatrix}^{-1} \begin{bmatrix} \alpha(\rho M_H + S_H)^{\mathcal{I}^k,\mathcal{I}^k} & 0 & L^{\mathcal{I}^k,:} \\ 0 & M_\partial & N \\ L^{:,\mathcal{I}^k} & N & 0 \end{bmatrix} \begin{bmatrix} \bar{v}^{\mathcal{I}^k} \\ \bar{u}^k \\ \bar{w}^k \end{bmatrix} = \begin{bmatrix} (\rho M_H + S_H)^{\mathcal{I}^k,\mathcal{I}^k} & 0 & 0 \\ 0 & S_B & 0 \\ 0 & 0 & S_B \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ M_\partial \bar{d} \\ 0 \end{bmatrix}.$$

We thus use the preconditioner

$$[\bar{\mathcal{R}}^k]^{-1} = \begin{bmatrix} (\rho M_H + S_H)^{\mathcal{I}^k,\mathcal{I}^k} & 0 & 0 \\ 0 & S_B & 0 \\ 0 & 0 & S_B \end{bmatrix}^{-1}. \tag{67}$$

Finally, we update the Lagrange multiplier $\bar{\lambda}^{\mathcal{A}^k}$ by solving

$$(\rho M_H + S_H)^{\mathcal{A}^k,\mathcal{A}^k} \bar{\lambda}^{\mathcal{A}^k} = L^{\mathcal{A}^k,:} \bar{w}^k,$$

where " $\mathcal{A}^k$ " denotes the active indices. (The derivation of this update is similar to the one leading to (59).

8.3 A $H^1(\Omega_v)$ control space on a regular grid

We have already discussed that the lack of a continuous PDAS algorithm for cases involving a $H^1(\Omega_v)$ control space do not seem to affect the performance of the preconditioner for the inverse ECG problem studied above. Now, we explore this issue further by considering the optimization problem

$$\min_{(v,u)\in H^1(\Omega_v)\times H^1(\Omega)} \left\{ \frac{1}{2}\|Tu - d\|^2_{L^2(\partial\Omega)} + \frac{1}{2}\alpha\|v\|^2_{H^1(\Omega_v)} \right\} \tag{68}$$

subject to

$$\int_\Omega \nabla\psi \cdot \nabla u \, dx + \int_\Omega \psi u \, dx = -\int_{\Omega_v} \nabla\psi \cdot \nabla v \, dx, \ \forall \psi \in H^1(\Omega), \tag{69}$$

$$v(x) \geq 0, \ x \in \Omega_v. \tag{70}$$

The domains Ω and Ω_v are defined as follows:

$$\Omega = (0,1) \times (0,1),$$
$$\Omega_v = \left(\frac{1}{4}, \frac{3}{4}\right) \times \left(\frac{1}{4}, \frac{3}{4}\right).$$

We will not present all the computational details, but instead focus on the iteration numbers for the preconditioned MINRES scheme applied to the KKT system associated with (68)-(70).

$N \backslash k$	0	1
64	237	209
128	273	227
256	297	277
512	334	275
1024	390	351

Table 6: The number of MINRES iterations required to solve the optimization problem (68)-(70). Note that k denotes the PDAS iteration number. For $k = 0$ there are *no* active constraints, whereas for $k = 1$ many constrains are active. Here, the stopping criterion was $\epsilon = 10^{-10}$, see (62), $\alpha = 0.01$, and the initial guess was set to $\bar{p}_0^k = 0$ for each PDAS iteration.

From Table 6 we conclude, at least for this problem, that there are no practical difficulties with combining our preconditioner with the PDAS algorithm. On the contrary, we observe a decrease in the number of MINRES iterations needed for $k = 1$, compared with the results obtained for $k = 0$. Note that, in the first PDAS iteration, i.e. $k = 0$, there are *no* active constraints, whereas for $k = 1$ many constrains are active. Hence, for this problem, the lack of a well defined extension operator $E^{\mathcal{I}^k}$, see (17)-(19), does not seem to introduce any severe difficulties. Nevertheless, further theoretical investigations are needed to develop a robust PDAS algorithm for PDE-constrained optimization problems with $H^1(\Omega_v)$ control spaces.

9 Conclusions

In this article we have analyzed the KKT systems arising in each iteration of the PDAS algorithm applied to PDE-constrained optimization problems with box constraints. More specifically, we have investigated whether the system

$$\mathcal{B}_\alpha^k p^k = b$$

can be solved efficiently with the MINRES method. Here, α is the Tikhonov regularization parameter, and $\mathcal{B}_\alpha^k$ denotes the indefinite Hermitian operator arising in each iteration of the PDAS scheme.

Our main theoretical result shows that the discretized operator $\mathcal{B}_\alpha^{k,h}$, associated with $\mathcal{B}_\alpha^k$, has a spectrum with a very limited number $N(\alpha)$ of isolated eigenvalues, whereas the remaining eigenvalues are contained in three bounded intervals:

$$\text{sp}(\mathcal{B}_\alpha^{k,h}) \subset [-b,-a] \cup [c\alpha, 2\alpha] \cup \{\lambda_1, \lambda_2, ..., \lambda_{N(\alpha)}\} \cup [a,b]. \tag{71}$$

For severely ill-posed problems $N(\alpha) = O(\ln(\alpha^{-1}))$. Theoretically, we therefore conclude that the MINRES algorithm will solve the KKT systems efficiently. Furthermore, since the spectral condition number $\kappa(\mathcal{B}_\alpha^{k,h})$ of $\mathcal{B}_\alpha^{k,h}$ is of order $O(\alpha^{-1})$, and the standard theory for the MINRES method states that $O(\kappa(\mathcal{B}_\alpha^{k,h}))$ iterations are required, we conclude that the classical analysis provides a pessimistic estimate for the needed workload.

In [11] it was established that the spectrum of the KKT system associated with (1)-(2), without inequality constraints, is on the form (71). From a technical point of view, the main challenge addressed in this paper was to prove that this property is inherited by the KKT system arising in each iteration of the PDAS method.

We presented a number of numerical experiments. In the first synthetic example, Example 1, we were interested in the growth of the iteration numbers with respect to both the regularization parameter α and the mesh parameter h. For the parameter α, we observed iteration counts almost of order

$$O(\ln(\alpha^{-1}))$$

as $\alpha \to 0$. Moreover, tables 1 and 2 show that the algorithm is robust with respect to the mesh parameter h. Theoretically, the spectral condition numbers of the KKT systems are bounded independently of any $h > 0$, and the slight increase we observed in practice is probably due to computational issues with the algebraic multigrid scheme.

In Section 8 we presented results for a real world problem. Namely, the *inverse problem of electrocardiography (ECG)* in which the unknown source is an ischemic region in the heart. Also for this problem, iteration counts approximately of order $O(\ln(\alpha^{-1}))$ were obtained. The numbers were, however, much higher than the iteration counts encountered in Example 1. This can be due to a number of reasons: The size of the domain, the unstructured grid, the noise in the data, or the form of the state equations. All these issues should be investigated properly in a separate paper.

Neither the inverse ECG problem, nor the synthetic example considered in Section 8.3, fulfill all the assumptions needed by our theoretical analysis. More specifically, these examples involve an H^1 control space, such that suitable extension operators, needed by the PDAS scheme, are not readily available. Nevertheless, our experiments revealed that solving the associated KKT systems, with many active constraints, did not require more MINRES iterations than solving unconstrained problems. Also, we obtained a rather

limited growth in the iteration numbers, as α decreased, for the real world application. In fact, we solved this problem in roughly 5 to 15 minutes, depending on the value of regularization parameter α. With optimized preconditioners, code optimization and a stronger CPU, it should be possible to reduce the computing time to less than 1 minute. For example, by changing the preconditioner (67) to

$$[\mathcal{R}^k]^{-1} = \begin{bmatrix} (\rho M_H + S_H)^{\mathcal{I}^k,\mathcal{I}^k} & 0 & 0 \\ 0 & N & 0 \\ 0 & 0 & N \end{bmatrix}^{-1}, \tag{72}$$

we get iteration counts as reported in Table 7. Clearly, substituting the stiffness matrix S_B in (67) with the matrix N, associated with the operator $-\nabla \cdot M\nabla$ on Ω_B, reduces the iteration counts and computing time significantly.

$k\backslash\alpha$	1	$10^{-1/2}$	10^{-1}	$10^{-3/2}$	10^{-2}
0	993	1528	2194	2661	3085
1	621	953	1224	1622	1715
2	191	444	693	817	948
Mean	602	975	1370	1700	1916
Wall Time	177s	285s	390s	471s	518s

Table 7: The number of MINRES iterations required to solve the optimization problem for Patient 1. These numbers were generated with the alternative preconditioner (72). Note that k denotes the PDAS iteration number. Here, the stopping criterion was $\epsilon = 10^{-6}$, see (62).

The overall conclusion of this paper is: By combining the MINRES method and the PDAS algorithm, some PDE constrained optimization problems arising in real world applications can be solved within reasonable time limits.

Acknowledgements

The authors would like to express their sincere gratitude to the FEniCS community. In particular we would like to thank Kent-André Mardal and Martin Sandve Alnæs for their contribution to the code development. We would also like to thank Kristina Hermann Haugaa, Andreas Abildgaard and Jan Gunnar Fjeld at Oslo University Hospital and Simula Research Laboratory for providing the clinical data used in this study.

A

We will discretize the update of the Lagrange multiplier in Example 1, see the discussion preceding (59). The generic update for this multiplier is given in (36) as

$$\lambda^{\mathcal{A}^k} = [B^{\mathcal{A}^k}]^* w^k, \tag{73}$$

where in each iteration of the PDAS method

$$B^{\mathcal{A}^k} = BE^{\mathcal{A}^k},$$

see (28). Furthermore, recall from (50) that

$$B = R_U^{-1}\widehat{B}.$$

It then follows from (28) that

$$\begin{aligned} B^{\mathcal{A}^k} &= R_U^{-1}\widehat{B}E^{\mathcal{A}^k} \\ &= R_U^{-1}\widehat{B}^{\mathcal{A}^k}, \end{aligned}$$

where

$$\widehat{B}^{\mathcal{A}^k} = \widehat{B}E^{\mathcal{A}^k}.$$

The update (73) involves the adjoint operator $[B^{\mathcal{A}^k}]^*$ of $B^{\mathcal{A}^k}$. According to a rather technical argument presented in [11],

$$[B^{\mathcal{A}^k}]^* = [R_{L^2(\mathcal{A}^k)}]^{-1}[\widehat{B}^{\mathcal{A}^k}]',$$

where the symbol "$'$" is used to denote dual operators and $R_{L^2(\mathcal{A}^k)}$ is the Riesz map of the space $L^2(\mathcal{A}^k)$, see (16). Hence, the continuous Lagrangian update in Example 1 is

$$\lambda^{\mathcal{A}^k} = [R_{L^2(\mathcal{A}^k)}]^{-1}[\widehat{B}^{\mathcal{A}^k}]' w^k,$$

or

$$R_{L^2(\mathcal{A}^k)}\lambda^{\mathcal{A}^k} = [\widehat{B}^{\mathcal{A}^k}]' w^k.$$

We again refer to [9] for further details about the discretization. Let the superscript notation "$\mathcal{A}^k$" and " : " denote the active indices and all the indices, respectively. The discretized update for the Lagrange multiplier then reads

$$M_v^{\mathcal{A}^k,\mathcal{A}^k}\bar{\lambda}^{\mathcal{A}^k} = M^{\mathcal{A}^k,:}\bar{w}^k.$$

References

[1] M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *ACTA NUMERICA*, 14:1–137, 2005.

[2] M. Bergounioux, K. Ito, and K. Kunisch. Primal-dual strategy for constrained optimal control problems. *SIAM Journal on Control and Optimization*, 37(4):1176–1194, 1999.

[3] L. T. Biegler, O. Ghattas, M. Heinkenschloss, and B. van Bloemen Waanders. *Large-Scale PDE-Constrained Optimization*. Springer, 2003.

[4] M. Engel and M. Griebel. A multigrid method for constrained optimal control problems. *Journal of Computational and Applied Mathematics*, 235(15):43684388, 2011.

[5] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, 2009.

[6] M. Hinze and M. Vierling. The semi-smooth Newton method for variationally discretized control constrained elliptic optimal control problems; implementation, convergence and globalization. *Optimization Methods and Software*, 27:933–950, 2012.

[7] K. Ito and K. Kunisch. *Lagrange multiplier approach to variational problems and applications*, volume 15 of *Advances in Design and Control*. Society for Industrial and Applied Mathematics (SIAM), Philidelphia, 2008.

[8] K.-A. Mardal and J. B. Haga. Block preconditioning of systems of PDEs. In Anders Logg, Kent-Andre Mardal, and Garth Wells, editors, *Automated Solution of Differential Equations*, pages 635–654. Springer, 2012.

[9] K.-A. Mardal and R. Winther. Preconditioning discretizations of systems of partial differential equations. *Numerical Linear Algebra with Applications*, 18(1):1–40, 2011.

[10] B. F. Nielsen and K.-A. Mardal. Efficient Preconditioners for Optimality Systems Arising in Connection with Inverse Problems. *SIAM Journal on Control and Optimization*, 48(8):5143–5177, October 2010.

[11] B. F. Nielsen and K.-A. Mardal. Analysis of the minimal residual method applied to ill-posed optimality systems. *SIAM Journal on Scientific Computing*, 35(2):785–814, 2013.

[12] C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12(4):617–629, 1975.

[13] A. J. Pullan, M. L. Buist, and L. K. Cheng. *Mathematically Modelling the Electrical Activity of the Heart: From Cell to Body Surface and Back*. World Scientific Publishing Company, 2005.

[14] M. Stoll and A. Wathen. Preconditioning for partial differential equation constrained optimization with control constraints. *Numerical Linear Algebra with Applications*, 19(1):53–71, 2012.

[15] F. Tröltzsch. *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society: Providence, Rhode Island, 2010.

[16] M. Ulbrich and S. Ulbrich. Primal-dual interior-point methods for PDE-constrained optimization. *Mathematical Programming*, 117(1-2):435–485, 2009.

3

Paper II - The split Bregman algorithm applied to PDE-constrained optimization problems with total variation regularization

This paper is submitted for publication.

The split Bregman algorithm applied to PDE-constrained optimization problems with total variation regularization

Ole Løseth Elvetun*and Bjørn Fredrik Nielsen†

November 8, 2014

Abstract

We derive an efficient solution method for ill-posed PDE-constrained optimization problems with total variation regularization. This regularization technique allows discontinuous solutions, which is desirable in many applications. Our approach is to adapt the split Bregman technique to handle such PDE-constrained optimization problems. This leads to an iterative scheme where we must solve a linear saddle point problem in each iteration. We prove that the spectra of the corresponding saddle point operators are almost contained in three bounded intervals, not containing zero, with a very limited number of isolated eigenvalues. Krylov subspace methods handle such operators very well and thus provide an efficient algorithm. In fact, we can guarantee that the number of iterations needed cannot grow faster than $O([\ln(\alpha^{-1})]^2)$ as $\alpha \to 0$, where α is a small regularization parameters. Moreover, in our numerical experiments we demonstrate that one can expect iteration numbers of order $O(\ln(\alpha^{-1}))$.

Keywords: Total Variation regularization, PDE-constrained optimization, Bregman algorithm, MINRES, KKT systems.

AMS subject classifications: 49K20, 35Q93, 65F22

1 Introduction

In the field of PDE-constrained optimization, sophisticated algorithms and increased computing power have made it possible to compute numerical

*Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, Norway. Email: ole.elvetun@nmbu.no

†Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, Norway; Simula Research Laboratory; Center for Cardiological Innovation, Oslo University Hospital. Email: bjorn.f.nielsen@nmbu.no

solutions of many advanced optimization problems. The use of Karush-Kuhn-Tucker (KKT) systems to solve such problems has become increasingly popular. These optimality systems are usually ill-posed, which leads to bad condition numbers for the discretized systems. Therefore, some kind of regularization technique must be invoked. The most popular method is the Tikhonov regularization technique, since this leads to linear optimality systems, provided that the state equation is also linear. In [18] the authors prove that a class of such saddle point systems can be solved efficiently by applying the Minimal Residual (MINRES) algorithm. More specifically, they prove that the eigenvalues of the discretized KKT system are almost contained in three bounded intervals. The number of isolated eigenvalues is only of order $O(\ln(\alpha^{-1}))$, where α is the regularization parameter. Krylov subspace methods are well suited to handle systems with such spectra.

It is known, however, that the use of a Tikhonov regularization term produces a smooth solution. In many inverse problems, the control parameter is often some physical property, like a heat source, density of a medium or an electrical potential. When we try to identify such quantities, it might be desirable to make a sharp separation between regions with different qualities of the physical property. In other words, we want "jumps" in the solution. Thus, one might argue that the smooth solutions produced with Tikhonov regularization are of limited value in such cases. The inverse problem of electrocardiography is a problem of this type, where one seeks to locate the ischemic[1] region of the heart. This can be achieved with a PDE-constrained optimization problem, where the control is the electrical potential in the myocardium, and the data is given in terms of ECG recordings. The ischemic area can be determined from the fact that the electrical potential is (approximately) piecewise constant, with different values in the ischemic and healthy regions. From an imaging point of view, it would be beneficial to properly separate these areas [27, 17].

In the field of image analysis, researchers have for decades been interested in optimization problems with such discontinuous solutions. In [22] the authors proposed the famous Rudin-Osher-Fatemi (ROF) model

$$\min_{v\in BV(\Omega)} \left\{ \frac{1}{2}\rho\|v-d\|^2_{L^2(\Omega)} + \int_\Omega |Dv|\ dx \right\}, \tag{1}$$

where the Banach space of functions with bounded variation is defined by

$$BV(\Omega) = \{v \in L^1(\Omega) : \int_\Omega |Dv|\ dx < \infty\}, \tag{2}$$

and the regularization term in (1) is defined by the distribution

$$\int_\Omega |Dv|\ dx = \sup \left\{ \int_\Omega v\ \text{div}\ p : p \in C_0^1(\Omega;\mathbb{R}^n);\quad |p|_\infty \le 1 \right\}. \tag{3}$$

[1] Ischemia is a precursor of heart infarction.

For elements in $W^{1,1}(\Omega)$, the distribution (3) is equal to the normal weak derivative, see [1].

The regularization term (3) is known as Total Variation (TV) regularization, and it allows for discontinuous solutions. This ability to include "jumps" in the solution has made it very popular for denoising pictures. Unfortunately, (1) is a very challenging problem to solve, since the TV term is highly non-linear and also non-differentiable. Nevertheless, due to the desirable denoising property, it has received much attention, and a large number of solution algorithms have been suggested, see e.g. [5, 25, 4].

The denoising case has been extended to include more sophisticated problems. In particular, the deblurring problem has been thoroughly analyzed. This problem can be written as

$$\min_{v\in BV(\Omega)} \left\{ \frac{1}{2}\rho\|\hat{K}v - d\|^2_{L^2(\Omega)} + \frac{1}{2}\kappa\|v\|^2_{L^2(\Omega)} + \int_\Omega |Dv| \; dx \right\}, \tag{4}$$

where $\hat{K} : BV(\Omega) \to L^2(\Omega)$ typically is a convolution operator, see e.g, [26]. The term $\frac{1}{2}\kappa\|v\|^2_{L^2(\Omega)}, 0 \leq \kappa << 1$, is added to guarantee uniqueness when K is non-injective [8]. This deblurring problem is the starting point for our PDE-constrained optimization formulation. Mathematically, an abstract form of a finite dimensional PDE-constrained optimization problem with TV regularization reads

$$\min_{(v_h,u_h)\in V_h\times U_h} \left\{ \frac{1}{2}\rho\|Tu_h - d_h\|^2_Z + \frac{1}{2}\kappa\|v_h\|^2_{L^2(\Omega)} + \int_\Omega |Dv_h| \; dx \right\}, \tag{5}$$

subject to

$$Au_h + Bv_h = 0, \tag{6}$$

where

1. $V_h = H^1_h(\Omega)$, i.e. P^1_h equipped with the H^1-norm, is the control space, $1 \leq \dim(V_h) < \infty$. The reason for this choice of norm will become clear when we discuss the preconditioner.

2. $U_h = U \cap P^n_h$ is the state space, $1 \leq \dim(U_h) < \infty$, and

3. $Z_h = Z \cap P^n_h$ is the observation space, $1 \leq \dim(Z_h) < \infty$.

From [8] it follows that this problem has a unique solution if $\kappa > 0$, or if $TA^{-1}B$ is injective. Here, U and Z are Hilbert spaces, P^n_h is a n-th order scalar FEM space and (6) is a discretized PDE. Furthermore, d_h is the given observation data, the domain $\Omega \subset \mathbb{R}^n$ is bounded, and $\rho > 1$ is the regularization parameter[2]. The operators A, B and T will be discussed properly in Section 3.

[2]Often the regularization parameter is placed in front of the regularization term and not in front of the data fidelity term. In the former approach, the values will then typically be $1/\rho$.

Note that we limit our study to discretized problems posed in terms of finite dimensional spaces. This simplifies the discussion of the TV-regularization and enables the use of the results published in [18].

For a specific kind of elliptic equation, the use of total variation regularization has been used to identify discontinuous coefficients. Basically, such problems have been solved with an augmented Lagrangian method, see e.g. [6, 9], or a level set method, see e.g. [14, 7].

The objective of this paper is to propose and analyze an efficient algorithm for solving the general problem (5)-(6). To succeed with this objective, we must not only guarantee an efficient iterative solution of the non-linear total variation term, but also for the inner systems that we will obtain in each iteration of the outer algorithm. This will be achieved by combining the analysis in [18] with a successful method for solving (4), namely the split Bregman method [10]. In more detail, we outline the paper as follows:

- Section 2 contains a brief introduction to the split Bregman algorithm.
- In Section 3 we show how the PDE-constrained optimization problem (5)-(6) can be modified in such a way that we can apply the split Bregman algorithm.
- In Section 4 we prove that the KKT systems that arise in each iteration of the split Bregman algorithm have a spectrum *almost* contained in three bounded intervals, with a very limited number of isolated eigenvalues. Hence, Krylov subspace algorithms will handle these systems very well. We will come back to the exact form of this spectrum in Section 4.
- Section 5 presents an alternative version of the split Bregman algorithm.
- In Section 6 we illuminate the theoretical results with some numerical experiments.
- Finally, the conclusions are presented in Section 7.

2 Brief overview of the split Bregman algorithm

The split Bregman method has its roots in the Bregman iteration, which is an algorithm for computing extrema of convex functionals [2]. Later, it was used in [19] as a new regularization procedure for inverse problems. In [10] the authors used this approach to find an effective solution method for L^1-regularization problems. In particular, they demonstrated why this method works well for total variation problems. The authors started by writing (4)

on the form

$$\min_{v_h,p_h\in V_h\times\mathbf{P}_h^0}\left\{\frac{1}{2}\rho\|Kv_h-d_h\|_{Z_h}^2+\frac{1}{2}\kappa\|v_h\|_{L_h^2(\Omega)}^2+\int_\Omega|p_h|\right\}, \tag{7}$$

subject to

$$\nabla v_h = p_h, \tag{8}$$

where $\mathbf{P}_h^0$ is a vector space of piecewise constant functions.

We will not go into details on how the split Bregman algorithm is derived, instead we refer to [10] and [3]. We would, however, like to highlight an interesting remark from [3]: Note that the problem (7)-(8) can be solved by sequentially solving the penalty formulation

$$\min_{v_h,p_h\in V_h\times\mathbf{P}_h^0}\left\{\begin{aligned}\frac{1}{2}\rho\|Kv_h-d_h\|_{Z_h}^2 \quad &+ \quad \frac{1}{2}\kappa\|v_h\|_{L_h^2(\Omega)}^2\\ &+ \quad \int_\Omega|p_h|+\frac{1}{2}\lambda^k\|\nabla v_h-p_h\|_{L_h^2(\Omega)}^2,\end{aligned}\right\}$$

where $\lambda^k\to\infty$. Unfortunately, such penalty methods are ineffective, and leads to numerical difficulties as λ^k grows large.

In the split Bregman algorithm, see Algorithm 1, we note that the parameter λ is fixed. Instead, it is the "data" that varies with the introduction of b^k, where b^k can be interpreted as the Lagrange multiplier estimate associated with (8). Hence, we obtain much better numerical stability. In [29] and [28] it is explained why the split Bregman scheme can be consider as an augmented Lagrangian method [11, 21].

Algorithm 1 The split Bregman algorithm for total variation regularization

1: Choose $v_h^0=0, p_h^0=0, b_h^0=0$
2: **for** k = 0, 1, 2,... **do**
3: $\quad v_h^{k+1}=\arg\min_{v_h\in V_h}\frac{1}{2}\rho\|Kv_h-d_h\|_{Z_h}^2+\frac{1}{2}\kappa\|v_h\|_{L_h^2(\Omega)}^2+\frac{1}{2}\lambda\|\nabla v_h-p_h^k+b_h^k\|_{L_h^2(\Omega)}^2$,
4: $\quad p_h^{k+1}=\arg\min_{p_h\in\mathbf{P}_h^0}\int_\Omega|p_h|+\frac{\lambda}{2}\|\nabla v_h^{k+1}-p_h+b_h^k\|_{L_h^2(\Omega)}^2$,
5: $\quad b_h^{k+1}=b_h^k+\nabla v_h^{k+1}-p_h^{k+1}$.
6: **end for**

Before we end this section, we would like to present one important theorem from [3]:

Theorem 2.1. *Assume that there exists at least one solution v_h^* of (7)-(8). Then the split Bregman algorithm satisfies*

$$\begin{aligned}\lim_{k\to\infty} \frac{1}{2}\rho\|Kv_h^k - d_h\|_{Z_h}^2 &+ \frac{1}{2}\kappa\|v_h^k\|_{L_h^2(\Omega)}^2 + \int_\Omega |\nabla v_h^k|\ dx \\ &= \frac{1}{2}\rho\|Kv_h^* - d_h\|_{Z_h}^2 + \frac{1}{2}\kappa\|v_h^*\|_{L_h^2(\Omega)}^2 + \int_\Omega |\nabla v_h^*|\ dx.\end{aligned}$$

If the solution v_h^* is unique, we also have

$$\lim_{k\to\infty} \|v_h^k - v_h^*\|_{L_h^2(\Omega)} = 0.$$

3 Split Bregman algorithm for PDE-constrained optimization problems

Recall that our main objective is to derive an efficient solution method for (5)-(6), i.e. for rather general PDE-constrained optimization problems subject to TV regularization. We will restrict our analysis to problems that satisfy the assumptions

$\mathcal{A}\mathbf{1}$: $A : U_h \to U_h'$ is bounded and linear.

$\mathcal{A}\mathbf{2}$: A^{-1} exists and is bounded.

$\mathcal{A}\mathbf{3}$: $B : V_h \to U_h'$ is bounded and linear.

$\mathcal{A}\mathbf{4}$: $T : U_h \to Z_h$ is bounded and linear.

Here, *bounded* should be interpreted as; having operator norm which is bounded independently of the mesh parameter h.

Due to assumption $\mathcal{A}\mathbf{2}$, we can write (6) on the form

$$u_h = -A^{-1}Bv_h. \tag{9}$$

Consequently, we might formulate the minimization problem (5)-(6) as

$$\min_{v\in V_h} \left\{ \frac{1}{2}\rho\|Kv_h - d_h\|_{Z_h}^2 + \frac{1}{2}\kappa\|v_h\|_{L_h^2(\Omega)}^2 + \int_\Omega |\nabla v_h|\ dx \right\} \tag{10}$$

where $K : V_h \to Z_h$ is defined by

$$K = -TA^{-1}B. \tag{11}$$

We observe that the minimization problem (10) is on the same form as (4). This motivates the use of the split Bregman algorithm. Unfortunately, however, the explicit computation of the operator K is not possible in practical applications; if (6) is a PDE, then the inverse of A is typically too expensive

to compute explicitly. This issue has been handled, in the case of Tikhonov regularization, by solving the associated KKT system. The purpose of this paper is to adapt the KKT approach to the framework of the split Bregman algorithm. As we will see below, this yields an efficient and practical solution method for PDE-constrained optimization problems subject to TV regularization.

We do this by applying Algorithm 1 to the minimization problem (10). Step 5 in Algorithm 1 is straightforward. Furthermore, Step 4 is, since[3] $\nabla v_h^{k+1}, p_h^k, b_h^k \in \mathbf{P}_h^0$, very cheap to solve by the shrinkage operator

$$p_{h,x_i}^{k+1}(x) = \text{shrink}\left(\nabla_{x_i} v_h^{k+1}(x) + b_{h,x_i}^k(x), \frac{1}{\lambda}\right), \tag{12}$$

where

$$\text{shrink}(a,b) = \frac{a}{|a|} * \max(a-b, 0),$$

see [10]. Hence, the challenge is to find the minimizer of Step 3. That is, we must solve the minimization problem

$$\min_{v_h \in V_h} \left\{ \frac{1}{2}\rho \|Kv_h - d_h\|_{Z_h}^2 + \frac{1}{2}\kappa \|v_h\|_{L_h^2(\Omega)}^2 + \frac{1}{2}\lambda \|\nabla v_h - p_h^k + b_h^k\|_{L_h^2(\Omega)}^2 \right\},$$

where d_h, p_h^k and b_h^k are given quantities. By combining this minimization problem with equations (9) and (11), we get the equivalent constrained minimization problem:

$$\min_{v_h, u_h \in V_h \times U_h} \left\{ \frac{1}{2}\rho \|Tu_h - d_h\|_{Z_h}^2 + \frac{1}{2}\kappa \|v_h\|_{L_h^2(\Omega)}^2 + \frac{1}{2}\lambda \|\nabla v_h - p_h^k + b_h^k\|_{L_h^2(\Omega)}^2 \right\} \tag{13}$$

subject to

$$Au_h + Bv_h = 0. \tag{14}$$

For the sake of simplicity, we want our optimality system to be as similar as possible to the optimality system analyzed in [18]. Thus, we need to scale the cost-functional in (13) such that we get

$$\min_{v_h, u_h \in V_h \times U_h} \left\{ \frac{1}{2}\|Tu_h - d_h\|_{Z_h}^2 + \frac{1}{2}\gamma \|v_h\|_{L_h^2(\Omega)}^2 + \frac{1}{2}\alpha \|\nabla v_h - p_h^k + b_h^k\|_{L_h^2(\Omega)}^2 \right\} \tag{15}$$

subject to (14), where

$$\alpha = \frac{\lambda}{\rho} \text{ and } \gamma = \frac{\kappa}{\rho}. \tag{16}$$

Next, we can introduce the Lagrangian associated with (14)-(15):

$$\begin{aligned} \mathcal{L}(v_h, u_h, w_h) &= \frac{1}{2}\|Tu_h - d_h\|_{Z_h}^2 + \frac{1}{2}\gamma \|v_h\|_{L_h^2(\Omega)}^2 \\ &+ \frac{1}{2}\alpha \|\nabla v_h - p_h^k + b_h^k\|_{L_h^2(\Omega)}^2 + \langle Au_h + Bv_h, w_h \rangle. \end{aligned}$$

[3] For higher order discretizations, the problem in Step 4 becomes more difficult to solve.

The first-order optimality conditions can be found by computing the derivatives of the Lagrangian with respect to v_h, u_h and w_h. These conditions can be expressed by the KKT system

$$\underbrace{\begin{bmatrix} -\alpha\Delta + \gamma E & 0 & B' \\ 0 & T'T & A' \\ B & A & 0 \end{bmatrix}}_{\widehat{\mathcal{A}}_\alpha} \begin{bmatrix} v_h \\ u_h \\ w_h \end{bmatrix} = \begin{bmatrix} -\alpha\nabla \cdot p_h^k + \alpha\nabla \cdot b_h^k \\ T'd_h \\ 0 \end{bmatrix}, \qquad (17)$$

where " $'$ " is used to denote dual operators, and $E : V_h \to V_h'$ is defined by

$$\langle Ev_h, \phi_h \rangle = (v_h, \phi_h)_{L_h^2(\Omega)}, \quad \phi_h \in V_h.$$

We have thus derived a new system of equations to be solved in Step 3 in Algorithm 1, which does not require the explicit inverse of A. Also note the form of the operator $-\Delta : V_h \to V_h'$ in the top left corner of the KKT system (17). In an infinite dimensional setting, this operator must be replaced with the more involved operator $D'D : BV(\Omega) \to BV(\Omega)'$, see [8] for a thorough discussion of this operator.[4] The operator $D'D$ is much more challenging to analyze, but it coincides with the operator $-\Delta$ in a finite dimensional setting, which follows from the fact that $Dv = \nabla v$ for all elements in $W^{1,1}(\Omega)$, see [1]. This concludes the discussion of Step 3 in Algorithm 1.

We might now formulate the full algorithm for solving the PDE-constrained optimization problem (5)-(6), see Algorithm 2.

Algorithm 2 The split Bregman algorithm for PDE-constrained optimization problems with TV regularization

1: Choose $v_h^0 = 0, p_h^0 = 0, b_h^0 = 0$
2: **for** k = 0, 1, 2,... **do**
3: Let (v_h^{k+1}, u_h^{k+1}, w_h^{k+1}) be the solution of (17).
4: $p_h^{k+1} = \arg\min_{p_h \in \mathbf{P}_h^0} \int_\Omega |p_h| + \frac{\lambda}{2} \|\nabla v_h^{k+1} - p_h + b_h^k\|_{L_h^2(\Omega)}^2$,
5: $b_h^{k+1} = b_h^k + \nabla v_h^{k+1} - p_h^{k+1}$.
6: **end for**

The efficiency of the split Bregman algorithm has been demonstrated earlier, see e.g. [10, 3]. Of the three inner steps of the **for**-loop in Algorithm 2, the update of b_h^k is obviously cheap, and the update of p_h^k is accomplished by the simple shrinkage operator (12). What remains, however, is to analyze the spectrum of the KKT system in (17), see Step 3 in Algorithm 2: The efficiency of the algorithm is highly dependent on how fast we can solve these KKT systems with, e.g., Krylov subspace solvers.

[4] In fact, it is possible to work with the dual space of BV with respect to the weak-* topology, which leads to the Laplacian instead of $D'D$ in a function space as well [12, 13].

4 Spectrum of the KKT system

In its current form, the operator $\widehat{\mathcal{A}}_\alpha$ in (17) is a mapping from the product space $V_h \times U_h \times U_h$ onto the dual space $V_h' \times U_h' \times U_h'$. Since this operator maps to the dual space, and not to the space itself, it is not possible to use the MINRES method directly. A remedy exists, however, in the form of Riesz maps. In this case, we must introduce the two Riesz maps

$$\begin{aligned} R_{V_h} &: V_h \to V_h', \\ R_{U_h} &: U_h \to U_h'. \end{aligned}$$

This enables us to use the MINRES algorithm, since the KKT system (17) can be written as

$$\begin{aligned} \underbrace{\begin{bmatrix} R_{V_h}^{-1} & 0 & 0 \\ 0 & R_{U_h}^{-1} & 0 \\ 0 & 0 & R_{U_h}^{-1} \end{bmatrix}}_{\mathcal{R}^{-1}} \underbrace{\begin{bmatrix} -\alpha\Delta + \gamma E & 0 & B' \\ 0 & T'T & A' \\ B & A & 0 \end{bmatrix}}_{\widehat{\mathcal{A}}_\alpha} \begin{bmatrix} v_h \\ u_h \\ w_h \end{bmatrix} \\ = \begin{bmatrix} R_{V_h}^{-1} & 0 & 0 \\ 0 & R_{U_h}^{-1} & 0 \\ 0 & 0 & R_{U_h}^{-1} \end{bmatrix} \begin{bmatrix} -\alpha\nabla \cdot p_h^k + \alpha\nabla \cdot b_h^k \\ T'd_h \\ 0 \end{bmatrix}, \end{aligned} \tag{18}$$

where

$$\mathcal{R}^{-1}\widehat{\mathcal{A}}_\alpha : V_h \times U_h \times U_h \to V_h \times U_h \times U_h.$$

The operator $\mathcal{R}^{-1}$ can be considered to be a preconditioner. See [16, 18] for a more thorough analysis.

We performed an experimental investigation that suggested the use of small values of α to obtain good convergence results for the outer split Bregman algorithm. That is, λ/ρ should be small, see (16). According to standard theory for Krylov subspace methods, the number of iterations needed by the MINRES algorithm is of the same order as the spectral condition number of the involved operator. In our case, this corresponds to iterations numbers of order $O(\alpha^{-1})$, when $\gamma = 0$. We will now show that this estimate is very pessimistic.

Since the case $\gamma = 0$ is the most challenging, and also the most interesting, we will for the rest of the analysis assume that this is the case, i.e. $\gamma = 0$. Let us first simplify the notation in (18), and write the operator $\mathcal{R}^{-1}\widehat{\mathcal{A}}_\alpha$ in the form

$$\mathcal{R}^{-1}\widehat{\mathcal{A}}_\alpha = \mathcal{A}_\alpha = \begin{bmatrix} \alpha Q & 0 & \tilde{B}^* \\ 0 & T^*T & \tilde{A}^* \\ \tilde{B} & \tilde{A} & 0 \end{bmatrix}, \tag{19}$$

where we have the following definitions:

- $Q = -R_{V_h}^{-1}\Delta : V_h \to V_h$,
- $\tilde{B} = R_{U_h}^{-1}B : V_h \to U_h$,
- $\tilde{A} = R_{U_h}^{-1}A : U_h \to U_h$,
- $T^*T = R_{U_h}^{-1}T'T : U_h \to U_h$.

In this new form, the operator $\mathcal{A}_\alpha$ in (19) is very similar to the operator analyzed in [18]. In fact, they analyzed the operator $\mathcal{B}_\alpha : V_h \times U_h \times U_h \to V_h \times U_h \times U_h$, defined as

$$\mathcal{B}_\alpha = \begin{bmatrix} \alpha I & 0 & \tilde{B}^* \\ 0 & T^*T & \tilde{A}^* \\ \tilde{B} & \tilde{A} & 0 \end{bmatrix}. \tag{20}$$

The main result in [18] is that the spectrum of $\mathcal{B}_\alpha$ is of the form

$$\mathrm{sp}(\mathcal{B}_\alpha) \subset [-b, -a] \cup [c\alpha, 2\alpha] \cup \{\tau_1, \tau_2, ..., \tau_{N(\alpha)}\} \cup [a, b],$$

where

$$N(\alpha) = O(\ln(\alpha^{-1})$$

and the constants a, b and $c > 0$ are independent of the parameter α. The analysis in [18] is roughly performed as follows:

- The negative eigenvalues are shown to be bounded away from zero, regardless of the size of regularization parameter $\alpha \geq 0$. That is, it even holds for $\alpha = 0$. Hence, the negative eigenvalues of $\mathcal{A}_\alpha$, defined in (19), are bounded away from zero: The argument in [18] can be adapted to the present situation in a straightforward manner.

- For the positive eigenvalues, the Courant-Fischer-Weyl min-max principle is used to show that the difference between the eigenvalues of $\mathcal{B}_0$ and $\mathcal{B}_\alpha$ is "small", where $\mathcal{B}_0$ denotes the operator $\mathcal{B}_\alpha$ with zero regularization $\alpha = 0$. More specifically, they prove that the difference between the eigenvalues of $\mathcal{B}_0$ and $\mathcal{B}_\alpha$, properly sorted, is less than the size of the regularization parameter $0 < \alpha \ll 1$. It is easy to verify that a similar property will hold for $\mathcal{A}_0$ and $\mathcal{A}_\alpha$. More specifically, the difference between the eigenvalues of $\mathcal{A}_0$ and $\mathcal{A}_\alpha$ is less than $\tilde{c}\alpha$, where $\tilde{c} = \|Q\|$.

- Finally, the analysis in [18] requires that

  $$\alpha(v_h, v_h)_{V_h} + (T^*Tu_h, u_h)_{U_h}$$

  must be coercive whenever

  $$\tilde{A}u_h + \tilde{B}v_h = 0.$$

It is proven in [18] that this property holds for the operator $\mathcal{B}_\alpha$. For the operator $\mathcal{A}_\alpha$, defined in (19), this analysis is more involved, and it will therefore be explored in detail here. More specifically, we must show, provided that suitable assumptions hold, that

$$\alpha(Qv_h, v_h)_{V_h} + (T^*Tu_h, u_h)_{U_h}$$

is coercive for all (v_h, u_h) satisfying

$$\tilde{A}u_h + \tilde{B}v_h = 0.$$

To further investigate the coercivity problem associated with (19), we introduce the notation

$$X_h = V_h \times U_h, \|x_h\|_{X_h} = \|(v_h, u_h)\|_{X_h} = \sqrt{\|v_h\|_{V_h}^2 + \|u_h\|_{U_h}^2},$$

$$M_\alpha = \begin{bmatrix} \alpha Q & 0 \\ 0 & T^*T \end{bmatrix} : X_h \to X_h, \tag{21}$$

$$N = \begin{bmatrix} \tilde{B} & \tilde{A} \end{bmatrix} : X_h \to U_h. \tag{22}$$

Since we work with finite dimensional spaces, we employ the control space V_h with the norm

$$\|\cdot\|_{V_h}^2 = \|\cdot\|_{L_h^2(\Omega)}^2 + |\cdot|_{H_h^1(\Omega)}^2, \tag{23}$$

i.e. $V_h = H_h^1(\Omega) \subset H^1(\Omega)$.

Note that, for the analysis presented below, we must assume that the operator B satisfies assumption $\mathcal{A}\mathbf{3}$ with the norm (23), i.e. that

$$B : V_h \to U_h'$$

is bounded, which along with assumptions $\mathcal{A}\mathbf{2}$ and $\mathcal{A}\mathbf{4}$ imply that

$$K_h = K = -TA^{-1}B = -T\tilde{A}^{-1}\tilde{B} : V_h \to Z_h$$

is bounded[5] (Bounded in the sense that the operator norm is bounded independently of h). We must also assume that the discrete solutions converge toward the correct solution as $h \to 0$:

$$\lim_{h\to 0} \|v_h - v\|_{H^1(\Omega)} = 0 \Rightarrow \lim_{h\to 0} \|K_h v_h - \hat{K}v\|_Z = 0, \tag{24}$$

where $\hat{K} : H^1(\Omega) \to Z$ denotes the associated mapping between the infinite dimensional spaces.

We are now ready to formulate the result concerning the coercivity issue for the operator $\mathcal{A}_\alpha$ defined in (19).

[5]Except for the presentation and discussion of Lemma 4.1, we simply write K instead of K_h.

Lemma 4.1. *Let M_α and N be defined as in (21) and (22), respectively. Assume that (24) holds and that $\hat{K}$ does not annihilate constants, i.e. the constant function $k \notin \mathcal{N}(\hat{K})$. Then there exists $\bar{h} > 0$ such that the operator M_α is coercive on the kernel of N, i.e. for $\alpha \in (0,1)$:*

$$(M_\alpha x_h, x_h)_{X_h} \geq c\alpha \|x_h\|^2_{X_h} \tag{25}$$

for all $h \in (0, \bar{h})$ and for all $x_h = (v_h, u_h) \in X_h$ satisfying

$$\tilde{A}u_h + \tilde{B}v_h = 0. \tag{26}$$

The constant c is independent of $h \in (0, \bar{h})$ and $\alpha > 0$.

Proof. We will first show that, if $\hat{K}$ does not annihilate constants, then there exist constants $\bar{h} > 0$ and $c \in (0,1)$, which is independent of $h \in (0, \bar{h})$, such that

$$\begin{aligned}(K_h v_h, K_h v_h)_Z \geq (c-1)(\nabla v_h, \nabla v_h)_{L^2(\Omega)} \\ + c(v_h, v_h)_{L^2(\Omega)}, \quad \forall v_h \in V_h,\, \forall h \in (0, \bar{h}).\end{aligned} \tag{27}$$

Thereafter, we will use this result to prove (25)-(26).

Assume that there do not exist $\bar{h} > 0$ and $c \in (0,1)$ such that (27) holds. We will show that this implies that the constant function k must belong to the null-space of $\hat{K}$. If (27) is not satisfied, then it follows that there exist a sequence

$$\lim_{i \to \infty} h_i = 0$$

and a sequence of functions

$$\{v_{h_i}\} \subset H^1(\Omega), \quad v_{h_i} \in H^1_{h_i}(\Omega), \quad \|v_{h_i}\|^2_{L^2(\Omega)} = 1,$$

such that

$$\begin{aligned}0 \leq (K_{h_i} v_{h_i}, K_{h_i} v_{h_i})_Z &< \left(\frac{1}{i} - 1\right) |v_{h_i}|^2_{H^1(\Omega)} + \frac{1}{i}(v_{h_i}, v_{h_i})_{L^2(\Omega)} \\ &= \left(\frac{1}{i} - 1\right) |v_{h_i}|^2_{H^1(\Omega)} + \frac{1}{i}.\end{aligned}$$

We may choose a sequence with the property $\|v_{h_i}\|^2_{L^2(\Omega)} = 1$ because the operator K_h is linear. Since $(1/i - 1) \to -1$ as $i \to \infty$, we can conclude that

$$|v_{h_i}|_{H^1(\Omega)} \to 0 \quad \text{as } i \to \infty, \tag{28}$$

$$\|K_{h_i} v_{h_i}\|_Z \to 0 \quad \text{as } i \to \infty. \tag{29}$$

We will now show that $\{v_{h_i}\}$ has a limit in $H^1(\Omega)$. Let

$$S = \left\{ s \in H^1(\Omega) : \int_\Omega s \, dx = 0 \right\}.$$

It is well known that $H^1(\Omega) = S \oplus \mathbb{R}$, i.e. every function in $H^1(\Omega)$ can be (uniquely) expressed as a sum of a function in S and a constant. Hence,

$$\begin{aligned} &v_{h_i} = s_{h_i} + r_i, \quad \text{where} \\ &s_{h_i} \in S, \\ &r_i \in \mathbb{R} \text{ is a constant.} \end{aligned}$$

From this splitting, we obtain

$$0 \leq |s_{h_i}|_{H^1(\Omega)} = |s_{h_i} + r_i|_{H^1(\Omega)} = |v_{h_i}|_{H^1(\Omega)} \to 0 \text{ as } i \to \infty, \tag{30}$$

see (28).

This enables us to use the Poincaré inequality to conclude that

$$0 \leq \|s_{h_i}\|_{L^2(\Omega)} \leq C|s_{h_i}|_{H^1(\Omega)} \to 0 \quad \text{as } i \to \infty,$$

i.e.

$$\|s_{h_i}\|_{L^2(\Omega)} \to 0 \quad \text{as } i \to \infty. \tag{31}$$

Furthermore, recall that $\|v_{h_i}\|^2_{L^2(\Omega)} = 1$ and that $\int_\Omega s_{h_i}\, dx = 0$. Thus, it follows that

$$\begin{aligned} 1 &= \|v_{h_i}\|^2_{L^2(\Omega)} \\ &= \|s_{h_i} + r_i\|^2_{L^2(\Omega)} \\ &= \|s_{h_i}\|^2_{L^2(\Omega)} + 2(s_{h_i}, r_i)_{L^2(\Omega)} + \|r_i\|^2_{L^2(\Omega)} \\ &= \|s_{h_i}\|^2_{L^2(\Omega)} + 2r_i \int_\Omega s_{h_i}\, dx + \|r_i\|^2_{L^2(\Omega)} \\ &= \|s_{h_i}\|^2_{L^2(\Omega)} + |\Omega|(r_i)^2, \end{aligned}$$

which yields

$$(r_i)^2 = \frac{1}{|\Omega|}\left(1 - \|s_{h_i}\|^2_{L^2(\Omega)}\right).$$

By using (31) we get

$$r_i = \frac{1}{\sqrt{|\Omega|}}\sqrt{1 - \|s_{h_i}\|^2_{L^2(\Omega)}} \to r^* = \frac{1}{\sqrt{|\Omega|}} \quad \text{as } i \to \infty.$$

We claim that also the sequence $\{v_{h_i}\}$ converges toward r^* in $H^1(\Omega)$:

$$v_{h_i} \to r^* = \frac{1}{\sqrt{|\Omega|}} \quad \text{in } H^1(\Omega).$$

This follows from the fact that $s_{h_i} = v_{h_i} - r_i$ and (30)-(31):

$$\begin{aligned}\|v_{h_i} - r^*\|_{H^1(\Omega)} &= \|v_{h_i} - r_i + r_i - r^*\|_{H^1(\Omega)} \\ &\leq \|v_{h_i} - r_i\|_{H^1(\Omega)} + \|r_i - r^*\|_{H^1(\Omega)} \\ &= \|s_{h_i}\|_{H^1(\Omega)} + \|r_i - r^*\|_{H^1(\Omega)} \\ &= \|s_{h_i}\|_{H^1(\Omega)} + \|r_i - r^*\|_{L^2(\Omega)} \xrightarrow{i\to\infty} 0.\end{aligned}$$

Here, we have used that $\{r_i\}$ is a sequence of constants and that r^* is a constant, which implies that $\|r_i - r^*\|_{H^1(\Omega)} = \|r_i - r^*\|_{L^2(\Omega)}$ and that

$$r_i \to r^* \text{ in } \mathbb{R} \Rightarrow \|r_i - r^*\|_{L^2(\Omega)} \to 0,$$

provided that Ω has finite measure.

Since $\{v_{h_i}\}$ converges toward r^* in $H^1(\Omega)$, we may employ assumption (24) to find that

$$\lim_{i\to\infty} \|K_{h_i} v_{h_i}\|_Z = \|\hat{K} r^*\|_Z.$$

By combining these observations with (29), we conclude that

$$r^* \in \mathcal{N}(\hat{K}).$$

To summarize, if (27) does not hold, then $\hat{K}$ annihilates constants. Conversely, if $\hat{K}$ does not annihilate constants, (27) must hold.

We are now ready to show that (25)-(26) does indeed hold. Note that (27) can be written on the form: There exists $c \in (0,1)$ such that

$$\begin{aligned}(\nabla v_h, \nabla v_h)_{L^2(\Omega)} \quad &+ \quad (K_h v_h, K_h v_h)_Z \\ &\geq \quad c[(\nabla v_h, \nabla v_h)_{L^2(\Omega)} + (v_h, v_h)_{L^2(\Omega)}], \\ &\quad\ \forall v_h \in V_h \text{ and } \forall h \in (0, \bar{h}).\end{aligned} \tag{32}$$

Assume that $x = (v_h, u_h) \in X_h$ satisfies the state equation, i.e.

$$\tilde{A} u_h + \tilde{B} v_h = 0.$$

Then,

$$u_h = -\tilde{A}^{-1} \tilde{B} v_h, \tag{33}$$

and since $\tilde{A}^{-1}\tilde{B}$ is assumed to be bounded independently of h,

$$\|u_h\|_{U_h} \leq \bar{c} \|v_h\|_{V_h}. \tag{34}$$

In addition,

$$T u_h = -T \tilde{A}^{-1} \tilde{B} v_h = K_h v_h. \tag{35}$$

Therefore, see (21), for $\alpha \in (0,1)$ and $h \in (0, \bar{h})$,

$$\begin{aligned}(M_\alpha x_h, x_h)_{X_h} &= \alpha(\nabla v_h, \nabla v_h)_{L^2_h(\Omega)} + (T^*Tu_h, u_h)_{U_h} \\ &\geq \alpha[(\nabla v_h, \nabla v_h)_{L^2_h(\Omega)} + (Tu_h, Tu_h)_{Z_h}] \\ &= \alpha[(\nabla v_h, \nabla v_h)_{L^2_h(\Omega)} + (K_h v_h, K_h v_h)_{Z_h}],\end{aligned}$$

where we have used (35). Next, by invoking (32) and (34) we can conclude that

$$\begin{aligned}(M_\alpha x_h, x_h)_{X_h} &\geq \alpha[(\nabla v_h, \nabla v_h)_{L^2_h(\Omega)} + (K_h v_h, K_h v_h)_{Z_h}], \\ &\geq \alpha c[(\nabla v_h, \nabla v_h)_{L^2_h(\Omega)} + (v_h, v_h)_{L^2_h(\Omega)}], \\ &\geq \alpha c[0.5(\nabla v_h, \nabla v_h)_{L^2_h(\Omega)} + 0.5(v_h, v_h)_{L^2_h(\Omega)} + 0.5\bar{c}^{-2}\|u_h\|^2_{U_h}], \\ &\geq \alpha c \min\{0.5, 0.5\bar{c}^{-2}\}\|(v_h, u_h)\|^2_{X_h}. \qquad (36)\end{aligned}$$

That is, M_α is coercive on the kernel of N, cf. (22). □

We will now use this lemma to establish the main result of this section. First, however, we note that:

Remark 1: From $\mathcal{A}$**1**-$\mathcal{A}$**3** it follows that the inf-sup condition holds, i.e.

$$\inf_{w_h \in U_h} \sup_{(v_h, u_h) \in V_h \times U_h} \frac{(\tilde{B}v_h, w_h)_{U_h} + (\tilde{A}u_h, w_h)_{U_h}}{\sqrt{\|v_h\|^2_{V_h} + \|u_h\|^2_{U_h}}\|w_h\|_{U_h}} \geq c > 0. \qquad (37)$$

Remark 2: Since we consider finite dimensional problems, there always exist positive constants $\tilde{c}$, $\tilde{C}$ such that

$$|\tau_i(\mathcal{A}_0)| \leq \tilde{c}e^{-\tilde{C}i}, \quad i = 1, 2, \ldots, n, \qquad (38)$$

where $\tau_i(\mathcal{A}_0)$ denotes the ith eigenvalue of $\mathcal{A}_0$ sorted in decreasing order according to their absolute value. Here, $\mathcal{A}_0$ is $\mathcal{A}_\alpha$ with $\alpha = 0$, i.e. without regularization. ($\mathcal{A}_\alpha$ is defined in (19)).

We consider ill-posed PDE-constrained optimization tasks. For such problems, $\tilde{c}e^{-\tilde{C}n}$ will typically be extremely small, i.e. much smaller than practical choices of the size of the regularization parameter.

Let us state the theorem:

Theorem 4.2. *Assume that all assumptions of Lemma 4.1 hold and that $h \in (0, \bar{h})$. Then there exist constants a, b, $c > 0$ such that, for $\alpha \in (0,1)$, the spectrum of $\mathcal{A}_\alpha$, defined in (19), satisfies*

$$\mathrm{sp}(\mathcal{A}_\alpha) \subset [-b, -a] \cup [c\alpha, 2\alpha] \cup \{\tau_1, \tau_2, ..., \tau_{N(\alpha)}\} \cup [a, b], \qquad (39)$$

where

$$N(\alpha) \leq \left\lceil \frac{\ln(\tilde{c}) - \ln(\alpha)}{\tilde{C}} \right\rceil = O(\ln(\alpha^{-1})).$$

Here $\tilde{c}$, $\tilde{C}$ are the constants in (38).

Proof. Since we consider finite dimensional problems, the theorem follows from Lemma 4.1 and the analysis presented in [18]. □

Since the spectrum of $\mathcal{A}_\alpha$ is of the form (39), we can conclude that the MINRES method will handle the KKT systems (18) very well. More precisely, the number of iterations needed by the MINRES scheme to solve (18) can not grow faster than $O([\ln(\alpha^{-1})]^2)$ as $\alpha \to 0$, see [18]. In fact, in practice, iterations counts of order $O(\ln(\alpha^{-1}))$ will in many situations occur, which is also explained in [18].

Note that, while the optimality system (5)-(6) requires that either $K = -TA^{-1}B : V_h \to Z_h$ is injective or that $\gamma > 0$ to obtain a unique solution, see [8], the inner MINRES algorithm only requires that the constant k does not belong to the null-space of $\hat{K}$.

5 Constrained split Bregman algorithm

The split Bregman algorithm we have analyzed is in [10] referred to as the *unconstrained* split Bregman method. For some applications, the related *constrained* split Bregman algorithm, also introduced in [10], produces better convergence rates. In order to discuss the latter method, we observe that the problem (4) can be formulated on the related, constrained form

$$\min_{v \in BV(\Omega)} \left\{ \frac{1}{2}\kappa \|v\|^2_{L^2(\Omega)} + \int_\Omega |p| \, dx \right\},$$

subject to

$$\begin{aligned} \hat{K}v &= d \quad \text{on } \Omega_{\text{observe}}, \\ Dv &= p \quad \text{on } \Omega. \end{aligned}$$

Here, Ω_{observe} is the domain on which the observation data d is defined. The constraints are "implicit" in the sense that they are not necessarily satisfied in each step of the split Bregman algorithm, see [10]. Instead, the scheme generates approximations which converge toward functions satisfying these constraints, and a natural stopping criterion is thus

$$\|\hat{K}v^k - d\|_Z < \text{TOL}.$$

Details about the constrained split Bregman algorithm associated with this problem can be found in [3].

It turns out that this constrained approach also can be applied to a PDE-constrained optimization problem, and an experimental investigation gave us better convergence results with this latter approach. We will therefore present the constrained split Bregman algorithm for discretized PDE-constrained optimization problems of the form

$$\min_{v_h \in V_h} \left\{ \frac{1}{2}\kappa \|v_h\|^2_{L^2_h(\Omega)} + \int_\Omega |p_h| \, dx \right\},$$

subject to

$$\begin{aligned} Au_h + Bv_h &= 0, \\ Tu_h &= d_h \quad \text{on } \Omega_{\text{observe}}, \\ Dv_h &= p_h \quad \text{on } \Omega. \end{aligned}$$

Note that the first constraint here is "explicit", i.e. it must be satisfied in each step of the algorithm. The latter two constraints are "implicit".

Recall the KKT system (17) that we derived in connection with Algorithm 2. For the constrained split Bregman method, we get the very similar optimality system

$$\underbrace{\begin{bmatrix} -\alpha\Delta + \gamma E & 0 & B' \\ 0 & T'T & A' \\ B & A & 0 \end{bmatrix}}_{\widehat{\mathcal{A}}_\alpha} \begin{bmatrix} v_h \\ u_h \\ w_h \end{bmatrix} = \begin{bmatrix} -\alpha\nabla \cdot p_h^k + \alpha\nabla \cdot b_h^k \\ T'd_h - T'c_h^k \\ 0 \end{bmatrix}, \tag{40}$$

where " $'$ " is used to denote dual operators, and $E : V_h \to V_h'$ is defined by

$$\langle Ev_h, \phi_h \rangle = (v_h, \phi_h)_{L_h^2(\Omega)}, \quad \phi_h \in V_h.$$

Compared with (17), only the term $-T'c_h^k$ has been added to the second row of the right hand side of (40). The operator $\widehat{\mathcal{A}}_\alpha$ on the left hand side is unchanged, and our analysis of the MINRES method, presented above, also applies to this KKT system. The associated algorithm is, of course, similar to Algorithm 2, see Algorithm 3.

Algorithm 3 The *constrained* split Bregman for PDE-constrained optimization problems with TV regularization

1: Choose $v_h^0 = 0, p_h^0 = 0, b_h^0 = 0$
2: **for** k = 0, 1, 2,... **do**
3: Let v_h^{k+1}, u_h^{k+1} and w_h^{k+1} be the solution of (40).
4: $p_h^{k+1} = \arg\min_{p_h \in \mathbf{P}_h^0} \int_\Omega |p_h| + \frac{\lambda}{2}\|\nabla v_h^{k+1} - p_h + b_h^k\|_{L_h^2(\Omega)}^2$,
5: $b_h^{k+1} = b_h^k + \nabla v_h^{k+1} - p_h^{k+1}$,
6: $c_h^{k+1} = c_h^k + Tu_h^{k+1} - d_h$.
7: **end for**

We observe that Algorithm 3 only requires one more simple update compared with Algorithm 2: The update for c_h^{k+1}. This extra computer effort is diminishingly small, and since we obtain better convergence results, we will present numerical experiments with the use of Algorithm 3 only.

6 Numerical experiments

6.1 Example 1

Let $\Omega = (0,1)\times(0,1)$. We consider the standard example in PDE-constrained optimization, but with TV regularization instead of Tikhonov regularization. That is,

$$\min_{(v_h,u_h)\in V_h\times U_h} \left\{ \frac{1}{2}\rho\|Tu_h - d_h\|^2_{L^2_h(\Omega)} + \int_\Omega |Dv_h| \right\}, \tag{41}$$

subject to

$$-\Delta u_h + u_h = v_h \text{ in } \Omega, \tag{42}$$
$$\nabla u_h \cdot n = 0 \text{ on } \partial\Omega, \tag{43}$$

where the control space V_h, the state space U_h and the observation space Z_h are

$$V_h = H^1_h(\Omega) = H^1(\Omega) \cap P^1_h, \tag{44}$$
$$U_h = H^1_h(\Omega), \tag{45}$$
$$Z_h = L^2_h(\Omega) = L^2(\Omega) \cap P^1_h, \tag{46}$$

respectively. Furthermore, the operator T is the embedding $T : H^1(\Omega) \hookrightarrow L^2(\Omega)$. Hence, assumption $\mathcal{A}\mathbf{4}$ is satisfied.

Recall that our objective is to solve this system with Algorithm 3. The main challenge is the efficient solution of the KKT systems (40). To derive this optimality system, we need the weak formulation of the boundary value problem (42)-(43).

Computational details. The weak formulation reads: Find $u_h \in U_h$ such that

$$\langle Au_h, \psi_h\rangle = -\langle Bv_h, \psi_h\rangle \quad \forall\psi_h \in U_h,$$

where

$$A : U_h \to U_h', \quad u_h \to \int_\Omega \nabla u_h \cdot \nabla\psi_h + u_h\psi_h \, dx, \ \forall\psi_h \in U_h, \tag{47}$$
$$B : V_h \to U_h', \quad v_h \to \int_\Omega v_h\psi_h \, dx, \forall\psi_h \in U_h. \tag{48}$$

From standard PDE theory, we find that A and A^{-1} have operator norms which are bounded independently of h. The boundedness of

$$B : V_h \to U_h',$$

where one employs the H^1-topology (23) on V_h, follows from the inequalities

$$\begin{aligned}\int_\Omega v_h\psi_h\,dx &\leq \|v_h\|_{L^2_h(\Omega)}\cdot\|\psi_h\|_{L^2_h(\Omega)}\\ &\leq \sqrt{\|v_h\|^2_{L^2_h(\Omega)}+|v_h|^2_{H^1_h(\Omega)}}\cdot\sqrt{\|\psi_h\|^2_{L^2_h(\Omega)}+|\psi_h|^2_{H^1_h(\Omega)}}\\ &= \|v_h\|_{V_h}\cdot\|\psi_h\|_{U_h}.\end{aligned}$$

We conclude that assumptions $\mathcal{A}\mathbf{1}$, $\mathcal{A}\mathbf{2}$ and $\mathcal{A}\mathbf{3}$ are satisfied.

The KKT system to be solved in Algorithm 3 now takes the form

$$\underbrace{\begin{bmatrix}R_{V_h}^{-1}&0&0\\0&R_{U_h}^{-1}&0\\0&0&R_{U_h}^{-1}\end{bmatrix}}_{\mathcal{R}^{-1}}\underbrace{\begin{bmatrix}-\alpha\Delta&0&B'\\0&T'T&A'\\B&A&0\end{bmatrix}}_{\widehat{\mathcal{A}}_\alpha}\begin{bmatrix}v_h\\u_h\\w_h\end{bmatrix}$$
$$=\begin{bmatrix}R_{V_h}^{-1}&0&0\\0&R_{U_h}^{-1}&0\\0&0&R_{U_h}^{-1}\end{bmatrix}\begin{bmatrix}-\alpha\nabla\cdot p_h^k+\alpha\nabla\cdot b_h^k\\T'd_h-T'c_h^k\\0\end{bmatrix}. \quad (49)$$

Recall that $\alpha=\lambda/\rho$, where ρ is the regularization parameter in (41) and λ is the parameter employed in the Bregman scheme, see the discussion of (13)-(16).

The discretization of the operator $\mathcal{R}$ in (49) is rather straightforward. Recall that the finite dimensional space V_h was equipped with the norm $\|\cdot\|_{H^1_h(\Omega)}$. Furthermore, since $U=H^1(\Omega)$ in this particular example, it follows that the discretization of both of the Riesz maps R_{V_h} and R_{U_h} yields the sum of the mass matrix M and stiffness matrix S.

For the operator $\widehat{\mathcal{A}}_\alpha$ in (49), the discretization is more challenging, but a general recipe can be found in [16]. The end result can be summarized as follows:

- A, defined in (47), yields the matrix $M+S$, which is the sum of the mass and stiffness matrices associated with the domain Ω.
- B, defined in (48), yields the mass matrix M.
- $-\Delta$ yields the stiffness matrix S.
- $T'T=R_{U_h}^{-1}T^*T$ yields the mass matrix M.
- The functions v_h, u_h, w_h, p_h^k, b_h^k, c_h^k and d_h yields the corresponding vectors $\bar{v}$, $\bar{u}$, $\bar{w}$, $\bar{p}^k$, $\bar{b}^k$, $\bar{c}^k$ and $\bar{d}$, respectively.

Hence, the matrix "version" of (49) is

$$\underbrace{\begin{bmatrix}(M+S)^{-1} & 0 & 0\\ 0 & (M+S)^{-1} & 0\\ 0 & 0 & (M+S)^{-1}\end{bmatrix}}_{\bar{\mathcal{R}}^{-1}}\underbrace{\begin{bmatrix}\alpha S & 0 & M\\ 0 & M & M+S\\ M & M+S & 0\end{bmatrix}}_{\bar{\mathcal{A}}_\alpha}\underbrace{\begin{bmatrix}\bar{v}^{k+1}\\ \bar{u}^{k+1}\\ \bar{w}^{k+1}\end{bmatrix}}_{\bar{q}^{k+1}}$$
$$=\begin{bmatrix}(M+S)^{-1} & 0 & 0\\ 0 & (M+S)^{-1} & 0\\ 0 & 0 & (M+S)^{-1}\end{bmatrix}\underbrace{\begin{bmatrix}-\alpha\nabla\cdot\bar{p}^k+\alpha\nabla\cdot\bar{b}^k\\ M\bar{d}-M\bar{c}^k\\ 0\end{bmatrix}}_{\bar{g}^k}. \tag{50}$$

The preconditioner thus reads

$$\begin{bmatrix}(M+S)^{-1} & 0 & 0\\ 0 & (M+S)^{-1} & 0\\ 0 & 0 & (M+S)^{-1}\end{bmatrix}, \tag{51}$$

and involves the inverse of the matrix $M+S$. This inverse is computed approximately by using algebraic multigrid (AMG). We discuss this in some more detail in the numerical setup.

Numerical setup.

- We wrote the code using cbc.block, which is a FEniCS-based Python implemented library for block operators. See [15] for details.
- The PyTrilinos package was used to compute an approximation of the preconditioner (51). We approximated the inverse using AMG with a symmetric Gauss-Seidel smoother with three smoothing sweeps. All tables containing iteration counts for the MINRES method were generated with this approximate inverse Riesz map. On the other hand, the eigenvalues of the KKT systems $[\bar{\mathcal{R}}]^{-1}\bar{\mathcal{A}}_\alpha$, see (50), were computed with an *exact* inverse $[\bar{\mathcal{R}}^k]^{-1}$ computed in Octave.
- To discretize the domain, we divided $\Omega=(0,1)\times(0,1)$ into $N\times N$ squares, and each of these squares were divided into two triangles.
- The MINRES iteration process was stopped as soon as

$$\frac{\|r_n^k\|}{\|r_0^k\|}=\left[\frac{(\ \bar{\mathcal{A}}_\alpha\bar{q}_n^k-\bar{g}^k,[\bar{\mathcal{R}}]^{-1}[\bar{\mathcal{A}}_\alpha\bar{q}_n^k-\bar{g}^k]\ )}{(\ \bar{\mathcal{A}}_\alpha\bar{q}_0^k-\bar{g}^k,[\bar{\mathcal{R}}]^{-1}[\bar{\mathcal{A}}_\alpha\bar{q}_0^k-\bar{g}^k]\ )}\right]^{1/2}<\epsilon. \tag{52}$$

  Here, ϵ is a small positive parameter. Note that the superindex k is the iteration index for the "outer" split Bregman method, while the subindex n is the iteration index for the "inner" MINRES algorithm (at each step of the split Bregman method).

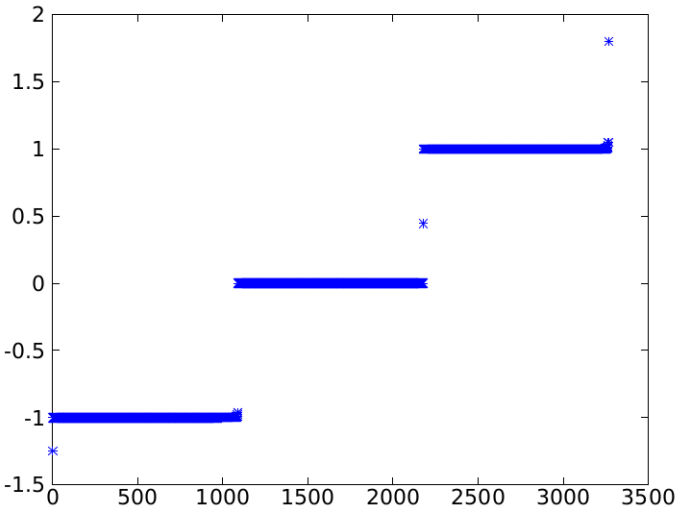


Figure 1: The eigenvalues of $[\bar{\mathcal{R}}^k]^{-1}\bar{\mathcal{A}}_\alpha$ in Example 1. Here, $\alpha = 0.0001$ and $N = 32$, i.e. $h = 1/32$. ($[\bar{\mathcal{R}}^k]^{-1}$ denotes the exact inverse of the preconditioner - not its AMG approximation).

- No noise was added to the input data d_h, see (41).

Results. We are now ready to solve the problem (41)-(43). The synthetic data d_h was produced by setting

$$v_h(x) = \begin{cases} -5 & \text{if } x_2 < 0.5, \\ 7 & \text{if } x_2 > 0.5, \end{cases} \tag{53}$$

and then we solved the boundary value problem (42)-(43) with (53) as input. The data d_h was thereafter set equal to the solution u_h throughout the entire domain $\Omega = (0,1) \times (0,1)$.

Theorem 4.2 states that the KKT system (18)-(19) arising in each iteration of the split Bregman iteration has a spectrum of the form (39). In Figure 1, we see a spectrum of such a KKT system, and it is clearly of the form (39). Hence, we should expect the MINRES algorithm to solve the problem efficiently.

Table 1 illuminates the theoretically established convergence behavior of the MINRES algorithm. As previously mentioned, in [18] the authors proved that the number of iterations can not grow faster than $O([\ln(\alpha^{-1})]^2)$, and showed why iteration growth of $O(\ln(\alpha^{-1}))$ often occur in practice. For $\epsilon = 10^{-6}$, see (52), and $N = 256$, we get the following estimate for the iteration growth

$$40.2 - 21.6 \log_{10}(\alpha),$$

where the coefficients are computed by the least squares method. The growth is very well modeled by this formula. Similarly, for $\epsilon = 10^{-10}$ and $N = 256$, we can model the growth by the formula

$$57.6 - 43.5 \log_{10}(\alpha).$$

N\α	1	.1	.01	.001	.0001
32	22	37	47	59	73
64	31	51	63	81	102
128	26	42	59	75	97
256	39	62	84	108	124

(a) Stopping criterion $\epsilon = 10^{-6}$.

N\α	1	.1	.01	.001	.0001
32	32	61	81	98	116
64	43	82	115	143	173
128	40	74	110	142	170
256	54	103	152	182	232

(b) Stopping criterion $\epsilon = 10^{-10}$.

Table 1: The average number of MINRES iterations required to solve the KKT systems arising in the first ten steps of the split Bregman algorithm in Example 1. The two panels display the iteration counts for two different choices of ϵ, see (52).

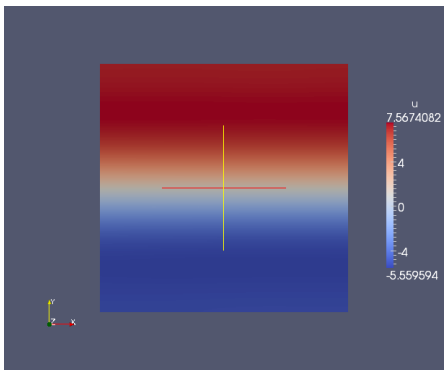
In Figure 2, four approximate solutions of the optimization problem (41)-(43) are displayed: After 10, 30, 50 and 70 Bregman iterations. From this figure, we observe that the jump is "found" after the first 30 iterations, cf. (53). The subsequent iterations merely "tightens" the jump and levels out the other parts of the solution. This behavior is similar to the one described for the image denoising algorithm in [10], where the authors also gave an explanation for why the split Bregman algorithm would quickly localize the jump(s).

Remark. As mentioned above, the problem (41)-(43), with Tikhonov regularization instead of TV regularization, has been analyzed by many scientists. In fact, for Tikhonov regularization a number of numerical schemes that are completely robust with respect to the size of the regularization parameter have been developed [24, 23, 20]: Even logarithmic growth in iterations counts is avoided. As far as the authors knows, it is not know whether these techniques can be adapted to the saddle point problem (49).

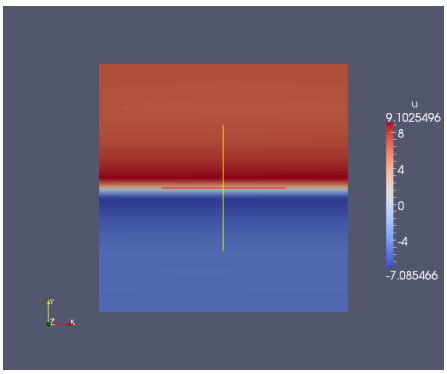
6.2 Example 2

We will now explore a more challenging problem. Let the domain Ω still be the unit square. Furthermore, define

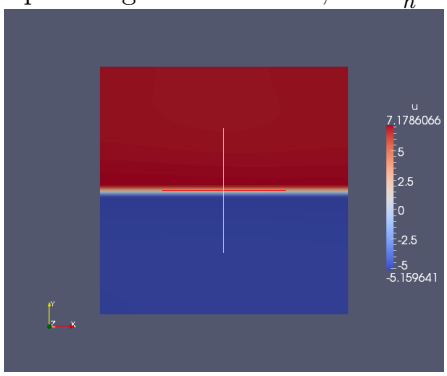
$$\tilde{\Omega} = (1/4, 3/4) \times (1/4, 3/4).$$



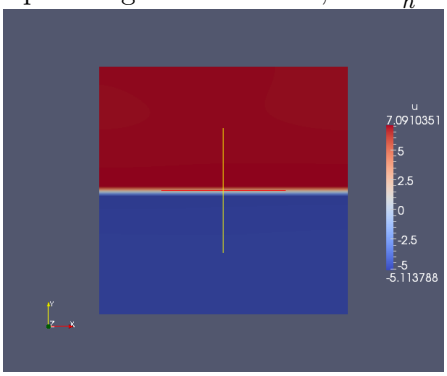
(a) Approximative inverse solution generated with 10 split Bregman iterations, i.e. v_h^{10}.



(b) Approximative inverse solution generated with 30 split Bregman iterations, i.e. v_h^{30}.



(c) Approximative inverse solution generated with 50 split Bregman iterations, i.e. v_h^{50}.



(d) Approximative inverse solution generated with 70 split Bregman iterations, i.e. v_h^{70}.

Figure 2: The solution of the problem (41)-(43). Here, $\epsilon = 10^{-6}$, $\alpha = 10^{-6}$ and $N = 128$ (*i.e.* $h = 1/128$).

The problem we want to study is

$$\min_{(v_h,u_h)\in V_h\times U_h}\left\{\frac{1}{2}\rho\|Tu_h-d_h\|^2_{L^2_h(\partial\Omega)}+\int_{\tilde{\Omega}}|Dv_h|\right\}, \tag{54}$$

subject to

$$\begin{aligned} -\Delta u_h+u_h &= \begin{cases} -v_h & \text{if } x\in\tilde{\Omega},\\ 0 & \text{if } x\in\Omega\setminus\tilde{\Omega}, \end{cases} && (55)\\ \nabla u_h\cdot n &= 0 \text{ on } \partial\Omega, && (56) \end{aligned}$$

where the control space V_h, the state space U_h and the observation space Z_h are

$$V_h = H^1_h(\tilde{\Omega}), \tag{57}$$
$$U_h = H^1_h(\Omega) = H^1(\Omega)\cap P^1_h, \tag{58}$$
$$Z_h = L^2_h(\partial\Omega) = L^2(\partial\Omega)\cap T(P^1_h), \tag{59}$$

respectively. Furthermore, the operator $T : H^1(\Omega)\to L^2(\partial\Omega)$ is the trace operator. Hence, assumption $\mathcal{A}\mathbf{4}$ is satisfied.

We observe two differences between examples 1 and 2. First, the control domain $\tilde{\Omega}$ is now a subdomain of the entire domain Ω, bounded strictly away from the boundary $\partial\Omega$. Secondly, the observation domain is reduced from the entire domain Ω to the boundary $\partial\Omega$.

In this model problem, V_h does not coincide with the control space defined in bullet point 1 in Section 1. Nevertheless, the proof of Lemma 4.1 can be adapted to the present situation in a straightforward manner, and Theorem 4.2 therefore also holds for this example.

Since the discretization of (54)-(56) is very similar to the discretization of (41)-(43), we do not enter into all the details. Instead, we only focus on the differences.

The weak formulation of the state equations (55)-(56) reads: Find $u\in U_h$ such that

$$\langle Au_h,\psi_h\rangle = -\langle Bv_h,\psi_h\rangle \quad \forall\psi_h\in U_h,$$

where the operator A is still defined as in (47). The operator B, however, is no longer as in (48), but is here defined by

$$B : V_h\to U_h', \quad v_h\to\int_{\tilde{\Omega}} v_h\psi_h\,dx, \forall\psi_h\in U_h, \tag{60}$$

where we can employ the norm

$$\|\cdot\|^2_{V_h} = \|\cdot\|^2_{L^2_h(\tilde{\Omega})}+|\cdot|^2_{H^1_h(\tilde{\Omega})}$$

on the control space V_h. From standard PDE theory, we can guarantee that A and A^{-1} are bounded, and the boundedness of B is verified in a manner very similar to the argument presented in connection with Example 1:

$$\begin{aligned}\int_{\tilde{\Omega}} v_h \psi_h \, dx &\leq \|v_h\|_{V_h} \cdot \|\psi_h\|_{H_h^1(\tilde{\Omega})} \\ &\leq \|v_h\|_{V_h} \cdot \|\psi_h\|_{U_h}\end{aligned}$$

because $\tilde{\Omega}$ is a subdomain of Ω. We conclude that assumptions $\mathcal{A}\mathbf{1}$, $\mathcal{A}\mathbf{2}$ and $\mathcal{A}\mathbf{3}$ are satisfied.

The new control domain $\tilde{\Omega}$ and the redefined operators B and T lead to some changes in the discretization of the optimality system (40), which must be solved repeatedly in Algorithm 3. These can be summarized as follows:

- B, defined in (60), yields the mass matrix $\tilde{M}$ associated with the subdomain $\tilde{\Omega}$.
- $-\Delta$ yields the stiffness matrix $\tilde{S}$ associated with the subdomain $\tilde{\Omega}$.
- $T'T = R_{U_h}^{-1} T^* T$ yields the "boundary" mass matrix M_∂.
- The Riesz map R_{V_h} now yields the sum of the mass matrix $\tilde{M}$ and stiffness matrix $\tilde{S}$.

All other operators are discretized in the same fashion as in Example 1. Hence, the matrix "version" of the optimality system in Algorithm 3, associated with (54)-(56), takes the form

$$\begin{aligned}&\underbrace{\begin{bmatrix}(\tilde{M}+\tilde{S})^{-1} & 0 & 0 \\ 0 & (M+S)^{-1} & 0 \\ 0 & 0 & (M+S)^{-1}\end{bmatrix}}_{\bar{\mathcal{R}}^{-1}} \underbrace{\begin{bmatrix}\alpha\tilde{S} & 0 & \tilde{M} \\ 0 & M_\partial & M+S \\ \tilde{M} & M+S & 0\end{bmatrix}}_{\bar{\mathcal{A}}_\alpha} \underbrace{\begin{bmatrix}\bar{v}^{k+1} \\ \bar{u}^{k+1} \\ \bar{w}^{k+1}\end{bmatrix}}_{\bar{q}^{k+1}} \\ &\quad = \begin{bmatrix}(\tilde{M}+\tilde{S})^{-1} & 0 & 0 \\ 0 & (M+S)^{-1} & 0 \\ 0 & 0 & (M+S)^{-1}\end{bmatrix} \underbrace{\begin{bmatrix}-\alpha\nabla\cdot\bar{p}^k + \alpha\nabla\cdot\bar{b}^k \\ M_\partial\bar{d} - M_\partial\bar{c}^k \\ 0\end{bmatrix}}_{\bar{g}^k}.\end{aligned} \tag{61}$$

The preconditioner thus reads

$$\begin{bmatrix}(\tilde{M}+\tilde{S})^{-1} & 0 & 0 \\ 0 & (M+S)^{-1} & 0 \\ 0 & 0 & (M+S)^{-1}\end{bmatrix}. \tag{62}$$

Results. The synthetic data d_h was produced in the same manner as in Example 1. We computed the synthetic data from the function $v_h \in V_h$,

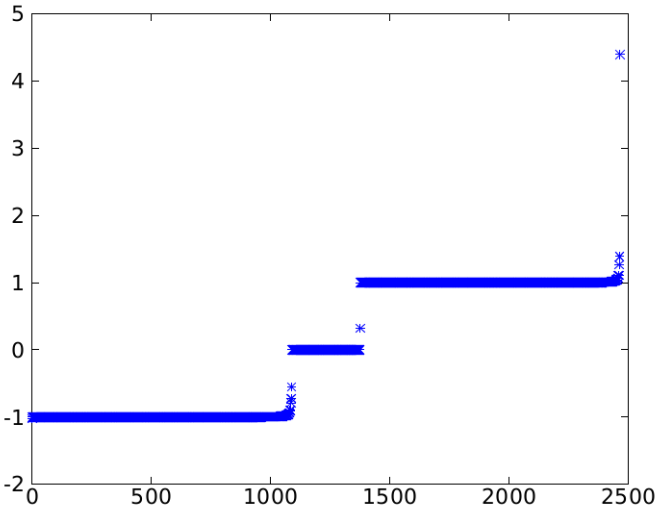


Figure 3: The eigenvalues of $[\bar{\mathcal{R}}^k]^{-1}\bar{\mathcal{A}}_\alpha$ in Example 2. Here, $\alpha = 0.0001$ and $N = 32$. ($[\bar{\mathcal{R}}^k]^{-1}$ denotes the exact inverse of the preconditioner - not its AMG approximation).

where

$$v_h(x) = \begin{cases} 5 & \text{if } x_1 < 0.5, \\ -5 & \text{if } x_1 > 0.5. \end{cases} \tag{63}$$

Note that the forward operator $K = -TA^{-1}B$ does not guarantee a unique solution of (54)-(56), since the trace operator is not injective, see [8]. Nevertheless, the forward operator K does not annihilate constants, and from Theorem 4.2 it then follows that the MINRES algorithm should handle the KKT systems, arising in each Bregman iteration, very well.

Figure 3 shows the spectrum of $[\bar{\mathcal{R}}^k]^{-1}\bar{\mathcal{A}}_\alpha$ for this example. This eigenvalue distribution is clearly on the form (39). Hence, in accordance with Theorem 4.2, we obtain such a spectrum even though $K = -TA^{-1}B$ is not injective (and $\kappa = 0$ in these computations).

Table 2 displays the iteration counts for Example 2. We see that the growth in the iteration numbers, as α decreases, is handled well by the MINRES algorithm. For example, for the case of $N = 256$ and $\epsilon = 10^{-6}$, the growth can be modeled by the formula

$$40.8 - 16.2\log_{10}(\alpha).$$

Similarly, for $N = 256$ and $\epsilon = 10^{-10}$, the least squares method gives us the formula

$$58.2 - 35.6\log_{10}(\alpha),$$

as the best logarithmic fit of iteration growth.

N\α	1	.1	.01	.001	.0001
32	29	44	49	55	63
64	34	48	58	67	82
128	36	52	59	69	84
256	41	60	71	84	110

(a) Stopping criterion $\epsilon = 10^{-6}$.

N\α	1	.1	.01	.001	.0001
32	41	65	82	100	109
64	47	76	104	126	154
128	50	84	112	144	169
256	57	95	131	163	201

(b) Stopping criterion $\epsilon = 10^{-10}$.

Table 2: The average number of MINRES iterations required to solve the KKT systems arising in the first ten steps of the split Bregman algorithm in Example 2. The two panels display the iteration counts for two different choices of ϵ, see (52).

The approximate solutions, seen in Figure 4, are very close to the "input solution" (63). We thus get very good approximations even though we can not guarantee a unique solution ($\kappa = 0$, see [8]).

In Figure 5 we show the inverse solution computed with Tikhonov regularization. Compared with the results obtained with total variation regularization, we observe that the latter produces far superior results for this problem. This is to be expected for problems with very sharp transition zones.

7 Conclusions

We have studied PDE-constrained optimization problems subject to TV regularization. The main purpose of this text was to adapt the split Bregman algorithm, frequently used in imaging analysis, to this kind of problems.

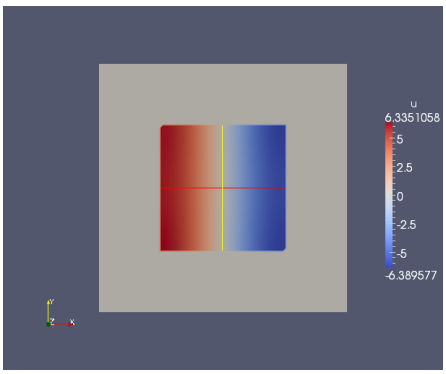
In each iteration of the split Bregman scheme, a large KKT system

$$\mathcal{A}_\alpha q = g \tag{64}$$

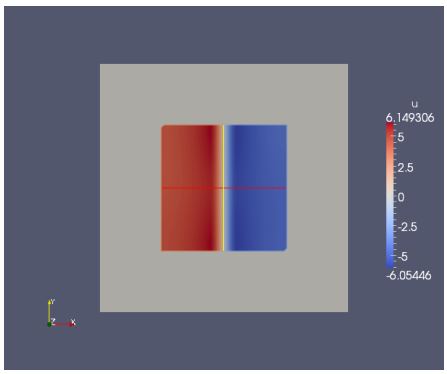
must be solved. Here, $0 < \alpha \ll 1$ is a regularization parameter, and the spectral condition number of $\mathcal{A}_\alpha$ tends to ∞ as $\alpha \to 0$. We investigated the performance of the MINRES algorithm applied to these indefinite systems. In particular, we analyzed the spectrum of $\mathcal{A}_\alpha$, and our main result shows that this spectrum is almost contained in three bounded intervals, with a small number of isolated eigenvalues. More specifically, we found that

$$\mathrm{sp}(\mathcal{A}_\alpha) \subset [-b, -a] \cup [c\alpha, 2\alpha] \cup \{\tau_1, \tau_2, ..., \tau_{N(\alpha)}\} \cup [a, b], \tag{65}$$

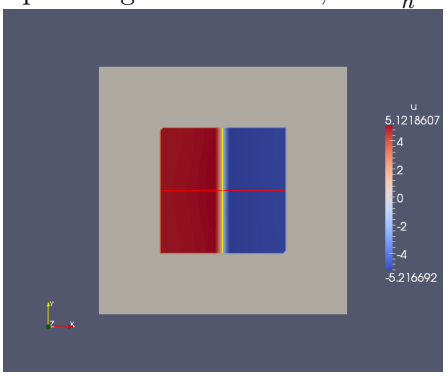
where $N(\alpha) = O(\ln(\alpha^{-1}))$. Krylov subspace solvers therefore handle (64) very well: The number of iterations required by the MINRES method can



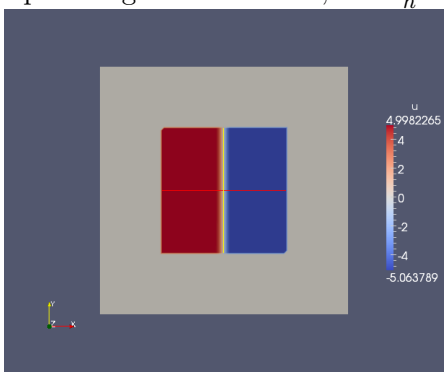
(a) Approximative inverse solution generated with 10 split Bregman iterations, i.e. v_h^{10}.



(b) Approximative inverse solution generated with 30 split Bregman iterations, i.e. v_h^{30}.



(c) Approximative inverse solution generated with 50 split Bregman iterations, i.e. v_h^{50}.



(d) Approximative inverse solution generated with 70 split Bregman iterations, i.e. v_h^{70}.

Figure 4: The solution of the problem (54)-(56). Here, $\epsilon = 10^{-6}$, $\alpha = 10^{-6}$ and $N = 128$ (*i.e.* $h = 1/128$).

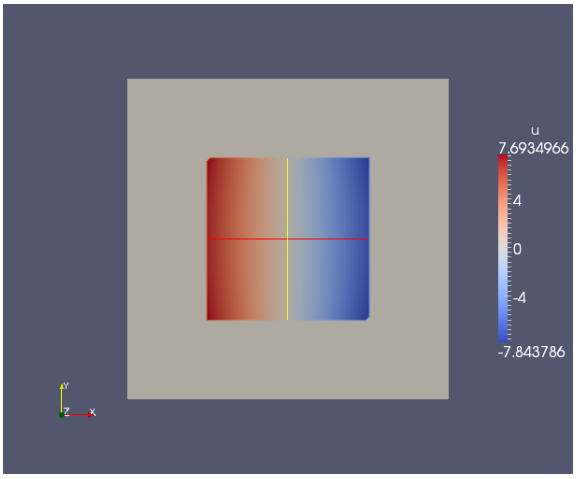


Figure 5: Inverse solution computed with standard Tikhonov regularization. The regularization parameter is 10^{-4}.

not grow faster than $O([\ln(\alpha^{-1})]^2)$ as $\alpha \to 0$, and in practice one will often encounter growth rates of order $O(\ln(\alpha^{-1}))$.

Our theoretical findings were illuminated by numerical experiments. In both examples we observed approximately logarithmic growth in iteration numbers as $\alpha \to 0$. This is in accordance with our theoretical results.

Acknowledgements

The authors would like to thank the FEniCS community for their work on the automatic solution of PDEs. We are also grateful for the comments provided by the referees, which significantly improved this article.

References

[1] Acar, R., Vogel, C.R.: Analysis of bounded variation penalty methods for ill-posed problems. Inverse Problems **10**, 1217–1229 (1994)

[2] Brègman, L.M.: A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. Zh. vychisl. Mat. mat. Fiz **7**, 620–631 (1967)

[3] Cai, J.F., Osher, S., Shen, Z.: Split Bregman methods and frame based image restoration. Multiscale modeling and simulation **8**, 337–369 (2009)

[4] Chambolle, A.: An algorithm for total variation minimization and application. Journal of Mathematical Imaging and Vision **20**, 89–97 (2004)

[5] Chan, T.F., Golub, G.H., Mulet, P.: A nonlinear primal-dual method for total variation-based image restoration. SIAM Journal on Scientific Computing **20**(6), 1964–1977 (1999)

[6] Chan, T.F., Tai, X.C.: Augmented Lagrangian and total variation methods for recovering discontinuous coefficients from elliptic equations. In: Computational Science for the 21st Century, pp. 597–607. John Wiley and Sons (1997)

[7] Chan, T.F., Tai, X.C.: Level set and total variation regularization for elliptic inverse problems with discontinuous coefficients. Journal of Computational Physics **193**, 40–66 (2004)

[8] Chavent, G., Kunisch, K.: Regularization of linear least squares problems by total bounded variation. ESAIM: Control, Optimisation and Calculus of Variations **2**, 359–376 (1997)

[9] Chen, Z., Zou, J.: An augmented Lagrangian method for identifying discontinuous parameters in elliptic systems. SIAM Journal on Control and Optimization **37**(3), 892–910 (1999)

[10] Goldstein, T., Osher, S.: The split Bregman method for L1-regularized problems. SIAM Journal on Imaging Sciences **2**, 323–343 (2009)

[11] Hestenes, M.R.: Multiplier and gradient methods. Journal of Optimization Theory and Applications **4**, 302–320 (1969)

[12] Hintermüller, M., Kunisch, K.: Totally bounded variation regularization as bilaterally constrained optimization problem. SIAM Journal on Applied Mathematics pp. 1311–1333 (2004)

[13] Hintermüller, M., Stadler, G.: A primal-dual algorithm for TV-based inf-convolution-type image restoration. SIAM Journal on Scientific Computing **28**, 1–23 (2006)

[14] Ito, K., Kunisch, K., Li, Z.: Level-set function approach to an inverse interface problem. Inverse Problems **17**(5), 1225–1242 (2001)

[15] Mardal, K.A., Haga, J.B.: Block preconditioning of systems of PDEs. In: A. Logg, K.A. Mardal, G. Wells (eds.) Automated Solution of Differential Equations, pp. 635–654. Springer (2012)

[16] Mardal, K.A., Winther, R.: Preconditioning discretizations of systems of partial differential equations. Numerical Linear Algebra with Applications **18**(1), 1–40 (2011)

[17] Nielsen, B.F., Lysaker, M., Grøttum, P.: Computing ischemic regions in the heart with the bidomain model - First steps towards validation. IEEE Transactions on Medical Imaging **32**, 1085–1096 (2013)

[18] Nielsen, B.F., Mardal, K.A.: Analysis of the minimal residual method applied to ill-posed optimality systems. SIAM Journal on Scientific Computing **35**(2), 785–814 (2013)

[19] Osher, S., Burger, M., Goldfarb, D., Xu, J., Yin, W.: An iterative regularization method for total variation-based image restoration. Multiscale Model. Simul **4**, 460–489 (2005)

[20] Pearson, J.W., Wathen, A.J.: A new approximation of the Schur complement in preconditioners for PDE-constrained optimization. Numerical Linear Algebra with Applications **19**(5), 816–829 (2012)

[21] Powell, M.J.D.: A method for nonlinear constraints in minimization problems. In Optimization ed. by R. Fletcher pp. 283–298 (1969)

[22] Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**, 259–268 (1992)

[23] Schöberl, J., Simon, R., Zulehner, W.: A robust multigrid method for elliptic optimal control problems. SIAM Journal on Numerical Analysis **49**(4), 1482–1503 (2011)

[24] Schöberl, J., Zulehner, W.: Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems. SIAM Journal on Matrix Analysis and Applications **29**(3), 752–773 (2007)

[25] Vogel, C.R., Oman, M.E.: Iterative methods for total variation denoising. SIAM Journal on Scientific Computing **17**, 227–238 (1996)

[26] Vogel, C.R., Oman, M.E.: Fast, robust total variation based reconstruction of noisy, blurred images. Image Processing, IEEE Transactions on **7**(6), 813–824 (1998)

[27] Wang, D., Kirby, R.M., MacLeod, R.S., Johnson, C.R.: Inverse electrocardiographic source localization of ischemia: An optimization framework and finite element solution. Journal of Computational Physics **250**, 403–424 (2013).

[28] Wu, C., Tai, X.C.: Augmented Lagrangian method, dual method and split-Bregman iterations for ROF, vectorial TV and higher order models. SIAM Journal on Imaging Sciences **3**(3), 300–339 (2010)

[29] Yin, W., Osher, S., Goldfarb, D., Darbon, J.: Bregman iterative algorithms for L1-minimization with applications to compressed sensing. SIAM Journal on Imaging Sciences **1**(1), 143–168 (2008)

4

Paper III - PDE-constrained optimization with local control and boundary observations: Robust preconditioners

This paper is submitted for publication.

PDE-constrained optimization with local control and boundary observations: Robust preconditioners

Ole Løseth Elvetun*and Bjørn Fredrik Nielsen†

December 5, 2014

Abstract

We consider PDE-constrained optimization problems with control functions defined on a subregion of the domain of the state equation. The main purpose of this paper is to define and analyze robust preconditioners for KKT systems associated with such optimization tasks. That is, preconditioners that lead to iteration bounds, for the MINRES scheme, that are independent of the regularization parameter α and the mesh size h.

Our analysis addresses elliptic control problems, subject to Tikhonov regularization, and covers cases with boundary observations only and locally defined control functions. A number of numerical experiments are presented.

Keywords: PDE-constrained optimization, preconditioning, minimal residual method.

AMS subject classifications: 49K20, 65F08, 65N21, 65F15.

1 Introduction

Parameter robust preconditioners for KKT systems arising in connection with PDE-constrained optimization have been successfully constructed [10, 11, 12, 13]. Nevertheless, these methods typically assume that observation data is available throughout the entire domain of the state equation, and that the control function also is defined on this domain. In [6] it is explained how one may handle problems with boundary observations only. The purpose of this text is to further explore this issue. More specifically, to investigate

*Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, Norway. Email: ole.elvetun@nmbu.no

†Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, Norway; Simula Research Laboratory; Center for Cardiological Innovation, Oslo University Hospital. Email: bjorn.f.nielsen@nmbu.no

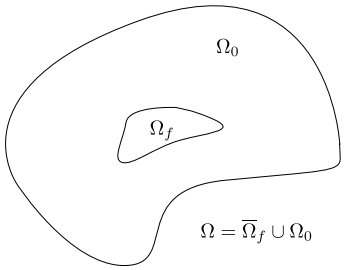


Figure 1: An example of a domain Ω with subdomains Ω_f and Ω_0.

how to design parameter robust preconditioners for problems with locally defined control functions and with boundary observations only.

Our work is motivated by the fact that many inverse problems, arising in the engineering sciences and in medical imaging, involve locally defined controls and limited observation data. This is, for example, the case for the inverse problem of electrocardiography (ECG).

2 Model problem

Consider the problem:

$$\min_{\mathbf{f},\,u} \left\{ \frac{1}{2} \left\| u - d \right\|^2_{L^2(\partial\Omega)} + \frac{1}{2}\alpha \left\| \mathbf{f} \right\|^2 \right\} \tag{1}$$

subject to

$$\Delta u - u = \begin{cases} \mathbf{f} & \text{in } \Omega_f, \\ 0 & \text{in } \Omega_0 = \Omega \setminus \overline{\Omega}_f, \end{cases} \tag{2}$$

$$\frac{\partial u}{\partial \mathbf{n}} = 0 \quad \text{on } \partial\Omega, \tag{3}$$

where $\mathbf{n}$ denotes the outward directed normal vector, of unit length, of $\partial\Omega$. We are thus aiming at using a L^2-boundary observation d, of u, to identify a source $\mathbf{f}$ defined on the subregion Ω_f of the domain Ω of the state equation. Note that Ω_0 represents the region $\Omega \setminus \overline{\Omega}_f$, see Figure 1. We assume that Ω_f and Ω are bounded and open domains, with Lipschitz boundaries, and that $\partial\Omega_f \cap \partial\Omega = \emptyset$.

Since the state u belongs to $H^1(\Omega)$, it is natural to seek a control $\mathbf{f}$ in the dual space $H^1(\Omega_f)'$, and the state equations (2)-(3) take the form

$$\int_\Omega \nabla u \cdot \nabla w + uw \, dx = -\langle \mathbf{f}, Rw \rangle \quad \forall w \in H^1(\Omega),$$

where $R : H^1(\Omega) \to H^1(\Omega_f)$ denotes the restriction operator, which, for the sake of simple notation, will be omitted.

Riesz' representation theorem implies that any $\mathbf{f} \in H^1(\Omega_f)'$ can be uniquely represented by a function $f \in H^1(\Omega_f)$. Hence, our optimization problem can be expressed as

$$\min_{f \in H^1(\Omega_f),\, u \in H^1(\Omega)} \left\{ \frac{1}{2} \|u - d\|^2_{L^2(\partial\Omega)} + \frac{1}{2}\alpha \|f\|^2_{H^1(\Omega_f)} \right\} \tag{4}$$

subject to

$$\int_\Omega \nabla u \cdot \nabla w + uw \, dx = -\int_{\Omega_f} \nabla f \cdot \nabla w + fw \, dx \quad \forall w \in H^1(\Omega). \tag{5}$$

Please note that (4)-(5) is similar to the inverse problem of electrocardiography, provided that the ST-shift in the transmembrane potential of the heart is used as the unknown source/control. But the inverse ECG problem involves conductivity tensors and the state equation does not contain any zero order terms.

3 Alternative formulation

We will now show that one can replace the state space $H^1(\Omega)$ in (4)-(5) with a function space consisting of functions satisfying, in a suitable weak sense, $\Delta\phi - \phi = 0$ in Ω_0. In Section 4 we employ this fact, and properly weighted Sobolev norms, to prove that the Brezzi conditions hold with α-independent constants. Thereafter, this insight is used to remove one of the unknowns from the KKT system and to develop parameter robust preconditioners.

From (5) it follows that the solution u of the state equation satisfies

$$(u, \psi)_{H^1(\Omega)} = 0 \quad \forall \psi \in S,$$

where

$$S = \left\{ \psi \in H^1(\Omega) \mid \psi|_{\overline{\Omega}_f} = 0 \right\},$$

i.e.

$$u \in U = S^\perp.$$

We will now briefly argue that $H^1(\Omega)$ in (4)-(5) can be replaced by U.

- Assume that u satisfies (5). Then, we know that $u \in U$ and, since U is a subspace of $H^1(\Omega)$, it follows that

$$(u, w)_{H^1(\Omega)} = -(f, w)_{H^1(\Omega_f)} \quad \forall w \in U. \tag{6}$$

- Let $w \in H^1(\Omega)$ be arbitrary and recall the orthogonal decomposition

$$w = q + q^\perp, \quad q \in S \text{ and } q^\perp \in U.$$

Assume that $u \in U$ satisfies (6). Then,

$$\begin{aligned}(u,w)_{H^1(\Omega)} &= (u,q)_{H^1(\Omega)} + (u,q^\perp)_{H^1(\Omega)} \\ &= (u,q^\perp)_{H^1(\Omega)} \\ &= -(f,q^\perp)_{H^1(\Omega_f)} \\ &= -(f,q^\perp)_{H^1(\Omega_f)} - (f,q)_{H^1(\Omega_f)} \\ &= -(f,w)_{H^1(\Omega_f)},\end{aligned}$$

where the second last equality follows from the fact that $q \in S$, i.e. $q|_{\overline{\Omega}_f} = 0$. Since $w \in H^1(\Omega)$ was arbitrary, we conclude that: If $u \in U$ satisfies (6), then u also satisfies (5).

It follows that we may rephrase (4)-(5) as follows:

$$\min_{f \in H^1(\Omega_f),\, u \in U} \left\{ \frac{1}{2} \|u - d\|^2_{L^2(\partial\Omega)} + \frac{1}{2}\alpha \|f\|^2_{H^1(\Omega_f)} \right\} \tag{7}$$

subject to

$$\int_\Omega \nabla u \cdot \nabla w + uw \, dx = -\int_{\Omega_f} \nabla f \cdot \nabla w + fw \, dx \quad \forall w \in U. \tag{8}$$

3.1 Helmholtz-harmonic extensions

We will now explain why functions in U can be regarded as Helmholtz-harmonic extensions, to the entire domain Ω, of functions defined on Ω_f. Below it will become evident why we use the term "Helmholtz-harmonic".

Let $\phi \in U = S^\perp$ be arbitrary. Then,

$$(\phi, \psi)_{H^1(\Omega)} = 0 \quad \forall \psi \in S,$$

or, since all $\psi \in S$ satisfy $\psi|_{\overline{\Omega}_f} = 0$,

$$\int_{\Omega_0} \nabla\phi \cdot \nabla\psi + \phi\psi \, dx = 0 \quad \forall \psi \in S,$$

where we recall that $\Omega_0 = \Omega \setminus \overline{\Omega}_f$, see Figure 1. The functions in S may be regarded as zero-extensions of functions belonging to $\{q \in H^1(\Omega_0) \mid q|_{\partial\Omega_f} = 0\}$. Thus, provided that the boundaries of Ω_f and Ω are Lipschitz, we may conclude that

$$\tilde{\phi} = \phi|_{\Omega_0}$$

is the weak solution of

$$\Delta\tilde{\phi} - \tilde{\phi} = 0 \quad \text{in } \Omega_0, \tag{9}$$

$$\tilde{\phi} = \phi \quad \text{on } \partial\Omega_f, \tag{10}$$

$$\frac{\partial\tilde{\phi}}{\partial\mathbf{n}} = 0 \quad \text{on } \partial\Omega. \tag{11}$$

We therefore refer to the functions in U as Helmholtz-harmonic on Ω_0.

Standard stability estimates and the trace theorem imply that

$$\|\phi\|_{H^1(\Omega_0)} = \|\tilde{\phi}\|_{H^1(\Omega_0)} \leq c\|\phi\|_{H^{1/2}(\partial\Omega_f)} \leq C\|\phi\|_{H^1(\Omega_f)}. \tag{12}$$

Throughout this text, c and C denote generic positive constants that are independent of the regularization parameter α and the grid parameter h.

Lemma 3.1. *There exists a positive constant c such that*

$$\|\phi\|_{H^1(\Omega_f)} \leq \|\phi\|_{H^1(\Omega)} \leq c\|\phi\|_{H^1(\Omega_f)} \quad \forall \phi \in U.$$

Proof. The first inequality follows from the assumption that $\Omega_f \subset \Omega$. The second inequality is a consequence of (12). □

4 KKT system

The Lagrangian associated with (7)-(8) reads

$$\begin{aligned} \mathcal{L}(f,u,w) &= \left\{ \frac{1}{2}\|u-d\|^2_{L^2(\partial\Omega)} + \frac{1}{2}\alpha\|f\|^2_{H^1(\Omega_f)} \right\} \\ &\quad + (f,w)_{H^1(\Omega_f)} + (u,w)_{H^1(\Omega)}, \end{aligned}$$

with $f \in H^1(\Omega_f)$, $u \in U$ and $w \in U$. And, from the first order optimality conditions

$$\frac{\partial \mathcal{L}}{\partial f} = 0, \; \frac{\partial \mathcal{L}}{\partial u} = 0, \; \frac{\partial \mathcal{L}}{\partial w} = 0,$$

we obtain the optimality system: Determine $(f,u,w) \in H^1(\Omega_f) \times U \times U$ such that

$$\begin{aligned} \alpha(f,\psi)_{H^1(\Omega_f)} + (\psi,w)_{H^1(\Omega_f)} &= 0 \quad \forall \psi \in H^1(\Omega_f), & (13)\\ (u-d,\phi)_{L^2(\partial\Omega)} + (\phi,w)_{H^1(\Omega)} &= 0 \quad \forall \phi \in U, & (14)\\ (f,\varphi)_{H^1(\Omega_f)} + (u,\varphi)_{H^1(\Omega)} &= 0 \quad \forall \varphi \in U. & (15) \end{aligned}$$

This system can be written on the form

$$\underbrace{\begin{bmatrix} \alpha\tilde{A}_f & 0 & A_f' \\ 0 & M_\partial & A' \\ A_f & A & 0 \end{bmatrix}}_{\mathcal{A}_\alpha^{3\times 3}} \begin{bmatrix} f \\ u \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ \tilde{M}_\partial d \\ 0 \end{bmatrix}, \tag{16}$$

where

$$A : U \to U', \qquad u \mapsto (u, \cdot)_{H^1(\Omega)}, \tag{17}$$
$$A_f : H^1(\Omega_f) \to U', \qquad f \mapsto (f, \cdot)_{H^1(\Omega_f)}, \tag{18}$$
$$\tilde{A}_f : H^1(\Omega_f) \to H^1(\Omega_f)', \qquad f \mapsto (f, \cdot)_{H^1(\Omega_f)}, \tag{19}$$
$$M_\partial : U \to U', \qquad u \mapsto (u, \cdot)_{L^2(\partial\Omega)}, \tag{20}$$
$$\tilde{M}_\partial : L^2(\partial\Omega) \to U', \qquad d \mapsto (d, \cdot)_{L^2(\partial\Omega)}. \tag{21}$$

The notation "$'$" is used to denote dual operators and dual spaces.

4.1 Weighted norms

For $\alpha > 0$, standard techniques can be employed to show that the Brezzi conditions hold for (16). In the standard L^2- and H^1-norms, however, the constants in the Brezzi conditions depend on α: Typically, the constant appearing in the coercivity condition is of order $O(\alpha)$. Consequently, we can not easily obtain an α-robust preconditioner with these norms. To remedy this, we can follow the procedure in [12] and introduce weighted Hilbert spaces, which are constructed in such a manner that the constants appearing in the Brezzi conditions are independent of the regularization parameter α.

For the control, state and multiplier spaces, we will work with the weighted norms

$$\begin{aligned} \|f\|^2_{F_\alpha} &= \alpha \|f\|^2_{H^1(\Omega_f)}, \\ \|u\|^2_{U_\alpha} &= \alpha \|u\|^2_{H^1(\Omega)} + \|u\|^2_{L^2(\partial\Omega)}, \\ \|w\|^2_{U_{\alpha^{-1}}} &= \frac{1}{\alpha} \|w\|^2_{H^1(\Omega)}. \end{aligned}$$

Note that we must have $\alpha > 0$ for these norms to make sense.

4.2 Inf-sup condition

The weighted norms give us the possibility to consider the operator $\mathcal{A}^{3\times 3}_\alpha$ in (16) as a mapping

$$\mathcal{A}^{3\times 3}_\alpha : F_\alpha \times U_\alpha \times U_{\alpha^{-1}} \to F'_\alpha \times U'_\alpha \times U'_{\alpha^{-1}}.$$

The analysis of saddle-point problems is standard and consists of three steps: Boundedness, coercivity on the kernel of the state equation, and the inf-sup condition. That the first two conditions are fulfilled, with α independent constants, follows from the results published in [9]. Their analysis will therefore be omitted. We are left with proving that the inf-sup condition holds, which we state in the following lemma:

Lemma 4.1. *There exists a constant $c > 0$, independent of $\alpha > 0$, such that*

$$\inf_{w\in U_{\alpha^{-1}}} \sup_{(f,u)\in F_\alpha\times U_\alpha} \frac{(f,w)_{H^1(\Omega_f)} + (u,w)_{H^1(\Omega)}}{\|(f,u)\|_{F_\alpha\times U_\alpha}\|w\|_{U_{\alpha^{-1}}}} \geq c.$$

Proof. Let $w \in U_{\alpha^{-1}} \setminus \{0\}$ be arbitrary. With $f = w|_{\Omega_f}$ and $u = 0$ we find that

$$\begin{aligned}
\sup_{(f,u)\in F_\alpha\times U_\alpha} \frac{(f,w)_{H^1(\Omega_f)} + (u,w)_{H^1(\Omega)}}{\|(f,u)\|_{F_\alpha\times U_\alpha}\|w\|_{U_{\alpha^{-1}}}} &\geq \frac{(w,w)_{H^1(\Omega_f)}}{\|w\|_{F_\alpha}\|w\|_{U_{\alpha^{-1}}}} \\
&= \frac{\|w\|^2_{H^1(\Omega_f)}}{\sqrt{\alpha}\|w\|_{H^1(\Omega_f)}\sqrt{\alpha^{-1}}\|w\|_{H^1(\Omega)}} \\
&\geq \frac{1}{\tilde{c}}\frac{\|w\|^2_{H^1(\Omega_f)}}{\|w\|_{H^1(\Omega_f)}\|w\|_{H^1(\Omega_f)}} \\
&= \frac{1}{\tilde{c}} = c,
\end{aligned}$$

where the last inequality follows from Lemma 3.1. □

We conclude that both $\|\mathcal{A}^{3\times3}_\alpha\|$ and $\|[\mathcal{A}^{3\times3}_\alpha]^{-1}\|$ are bounded independently of $\alpha > 0$.

5 Reducing the size of the KKT system

The main reason why the inf-sup condition in Lemma 4.1 holds, with an α independent constant, is the fact that $H^1(\Omega_f)$ and U are isomorphic. We will now see how this property also can be used to remove the Lagrange multiplier from the KKT system.

First, however, we will formalize the isomorphism between $H^1(\Omega_f)$ and U. Define the extension operator $E : H^1(\Omega_f) \to U$ as

$$E\phi = \begin{cases} \phi & \text{in } \Omega_f, \\ \tilde{\phi} & \text{in } \Omega_0, \end{cases} \tag{22}$$

where $\tilde{\phi}$ is the weak solution of (9)-(11). From standard theory for elliptic PDEs and Lemma 3.1, it follows that this operator is an isomorphism between $H^1(\Omega_f)$ and U.

Since both the state function, u, and the dual function, w, in the KKT system (13)-(15) belong to U, we may express them on the form

$$\begin{aligned} u &= E\hat{u}, \\ w &= E\hat{w}, \end{aligned}$$

where $\hat{u}, \hat{w} \in H^1(\Omega_f)$. Hence, equations (13)-(15) can be reformulated as

$$\begin{aligned} \alpha(f,\psi)_{H^1(\Omega_f)} + (\psi,\hat{w})_{H^1(\Omega_f)} &= 0 \quad \forall \psi \in H^1(\Omega_f), && (23)\\ (E\hat{u}-d, E\phi)_{L^2(\partial\Omega)} + (E\phi, E\hat{w})_{H^1(\Omega)} &= 0 \quad \forall \phi \in H^1(\Omega_f), && (24)\\ (f,\varphi)_{H^1(\Omega_f)} + (E\hat{u}, E\varphi)_{H^1(\Omega)} &= 0 \quad \forall \varphi \in H^1(\Omega_f). && (25) \end{aligned}$$

On this form, the relation between the control, f, and the dual, $\hat{w}$, becomes clear. In fact, from (23) it follows that

$$\hat{w} = -\alpha f.$$

Consequently, we can replace $\hat{w}$ with $-\alpha f$ in (24) to obtain the equations

$$\begin{aligned} -\frac{1}{\alpha}(E\hat{u}-d, E\phi)_{L^2(\partial\Omega)} + (Ef, E\phi)_{H^1(\Omega)} &= 0 \quad \forall \phi \in H^1(\Omega_f), && (26)\\ (E\hat{u}, E\varphi)_{H^1(\Omega)} + (f,\varphi)_{H^1(\Omega_f)} &= 0 \quad \forall \varphi \in H^1(\Omega_f). && (27) \end{aligned}$$

We can then write this system on the block form

$$\underbrace{\begin{bmatrix} -\frac{1}{\alpha}\mathsf{M}_\partial & \mathsf{A} \\ \mathsf{A} & \mathsf{A}_f \end{bmatrix}}_{\mathcal{A}_\alpha^{2\times 2}} \begin{bmatrix} \hat{u} \\ f \end{bmatrix} = \begin{bmatrix} -\frac{1}{\alpha}\tilde{\mathsf{M}}_\partial d \\ 0 \end{bmatrix}, \tag{28}$$

where

$$\begin{aligned} \mathsf{M}_\partial &: H^1(\Omega_f) \to H^1(\Omega_f)', && \hat{u} \mapsto (E\hat{u}, E\,\cdot\,)_{L^2(\partial\Omega)}, && (29)\\ \tilde{\mathsf{M}}_\partial &: L^2(\partial\Omega) \to H^1(\Omega_f)', && d \mapsto (d, E\,\cdot\,)_{L^2(\partial\Omega)}, && (30)\\ \mathsf{A} &: H^1(\Omega_f) \to H^1(\Omega_f)', && \hat{u} \mapsto (E\hat{u}, E\,\cdot\,)_{H^1(\Omega)}, && (31)\\ \mathsf{A}_f &: H^1(\Omega_f) \to H^1(\Omega_f)', && f \mapsto (f, \cdot\,)_{H^1(\Omega_f)}. && (32) \end{aligned}$$

We conclude that (28) has a unique solution since (16) has a unique solution and the extension operator $E : H^1(\Omega_f) \to U$ is isomorphic. (Also for other KKT-systems, arising in connection with PDE-constrained optimization, it is sometimes possible to reduce the problem to a 2×2 block system, see e.g. [13, 4].)

6 Analysis of the reduced system

For the original system (16) we have concluded that, in properly weighted Hilbert spaces, we have Brezzi constants which are independent of the regularization parameter α. It is not self-evident that similar properties can be proved for the reduced system (28).

The form of (28), and the fact that $Eu \in U$, motivate us to introduce the weighted norm

$$\|u\|^2_{U_{1+\alpha^{-1}}} = \|Eu\|^2_{H^1(\Omega)} + \frac{1}{\alpha}\|Eu\|^2_{L^2(\partial\Omega)},\ u \in H^1(\Omega_f),$$

for the state function. Note that we, for the sake of simplicity, no longer use the notation $\hat{u}$ for functions in $H^1(\Omega_f)$. To further increase readability, we define the product space

$$V_\alpha = U_{1+\alpha^{-1}} \times H^1(\Omega_f). \tag{33}$$

We can now define the bilinear form

$$a(\,\cdot\,;\,\cdot\,) : V_\alpha \times V_\alpha \to \mathbb{R},$$

associated with (28), as

$$\begin{aligned} a(u, f; \phi, \varphi) = &- \frac{1}{\alpha}(Eu, E\phi)_{L^2(\partial\Omega)} + (Ef, E\phi)_{H^1(\Omega)} \\ &+ (Eu, E\varphi)_{H^1(\Omega)} + (f, \varphi)_{H^1(\Omega_f)}. \end{aligned}$$

According to Babuška theory [1], (28) is well-posed if and only if $a(\,\cdot\,,\,\cdot\,)$ is continuous and weakly coercive. Since we in Section 5 concluded that (28) is well-posed, it follows that $a(\,\cdot\,,\,\cdot\,)$ fulfills these two conditions. To obtain an α-robust preconditioner, however, we must show that the constants appearing in the continuity and coercivity bounds are independent of α, provided that proper weighted norms are applied.

Lemma 6.1. *There exists a constant $C_1 > 0$, independent of $\alpha > 0$, such that*

$$|a(u, f; \phi, \varphi)| \leq C_1 \|(u, f)\|_{V_\alpha} \|(\phi, \varphi)\|_{V_\alpha},$$

where V_α is defined in (33).

Proof. Cauchy-Schwartz' inequality implies that

$$\begin{aligned} |a(u, f; \phi, \varphi)| &\leq \frac{1}{\sqrt{\alpha}} \|Eu\|_{L^2(\partial\Omega)} \frac{1}{\sqrt{\alpha}} \|E\phi\|_{L^2(\partial\Omega)} \\ &\quad + \|Ef\|_{H^1(\Omega)} \|E\phi\|_{H^1(\Omega)} \\ &\quad + \|Eu\|_{H^1(\Omega)} \|E\varphi\|_{H^1(\Omega)} \\ &\quad + \|f\|_{H^1(\Omega_f)} \|\varphi\|_{H^1(\Omega_f)} \\ &\leq \tilde{c} \left[\frac{1}{\sqrt{\alpha}} \|Eu\|_{L^2(\partial\Omega)} \frac{1}{\sqrt{\alpha}} \|E\phi\|_{L^2(\partial\Omega)} \right. \\ &\quad + \|f\|_{H^1(\Omega_f)} \|E\phi\|_{H^1(\Omega)} \\ &\quad + \|Eu\|_{H^1(\Omega)} \|\varphi\|_{H^1(\Omega_f)} \\ &\quad \left. + \|f\|_{H^1(\Omega_f)} \|\varphi\|_{H^1(\Omega_f)} \right] \\ &\leq 4\tilde{c} \|(u, f)\|_{V_\alpha} \|(\phi, \varphi)\|_{V_\alpha}, \end{aligned}$$

where $\tilde{c} = \max\{\|E\|, 1\}$. The result follows with $C_1 = 4\tilde{c}$. □

The weak coercivity of the bilinear operator $a(\,\cdot\,,\,\cdot\,)$ is defined in terms of two inf-sup conditions. Babuška theory asserts that the constants in the inf-sup conditions coincide when the system is well-posed, provided that only reflexive Banach spaces are involved. This constant will be independent of α, as the following lemma expresses.

Lemma 6.2 (Weak coercivity)**.** *There exists a constant $C_2 > 0$, independent of $\alpha > 0$, such that*

$$\inf_{(\phi,\varphi)\in V_\alpha} \sup_{(u,f)\in V_\alpha} \frac{a(u,f;\phi,\varphi)}{\|(u,f)\|_{V_\alpha}\|(\phi,\varphi)\|_{V_\alpha}} \geq C_2,$$

and

$$\inf_{(u,f)\in V_\alpha} \sup_{(\phi,\varphi)\in V_\alpha} \frac{a(u,f;\phi,\varphi)}{\|(u,f)\|_{V_\alpha}\|(\phi,\varphi)\|_{V_\alpha}} \geq C_2.$$

Proof. Let $(\phi,\varphi) \in V_\alpha \setminus \{(0,0)\}$ be arbitrary. With $u = -\phi$ and $f = \phi + \varphi$ we get

$$\begin{aligned}
&\sup_{(u,f)\in V_\alpha} \frac{a(u,f;\phi,\varphi)}{\|(u,f)\|_{V_\alpha}\|(\phi,\varphi)\|_{V_\alpha}} \\
&\qquad \geq \frac{\frac{1}{\alpha}\|E\phi\|^2_{L^2(\partial\Omega)} + \|E\phi\|^2_{H^1(\Omega)} + \|\varphi\|^2_{H^1(\Omega_f)} + (\phi,\varphi)_{H^1(\Omega_f)}}{\|(\phi,\phi+\varphi)\|_{V_\alpha}\|(\phi,\varphi)\|_{V_\alpha}} \\
&\qquad \geq \frac{\|\phi\|^2_{U_{1+\alpha^{-1}}} + \|\varphi\|^2_{H^1(\Omega_f)} - \|E\phi\|_{H^1(\Omega)}\|\varphi\|_{H^1(\Omega_f)}}{\|(\phi,\phi+\varphi)\|_{V_\alpha}\|(\phi,\varphi)\|_{V_\alpha}} \\
&\qquad \geq \frac{\|\phi\|^2_{U_{1+\alpha^{-1}}} + \|\varphi\|^2_{H^1(\Omega_f)} - \frac{1}{2}\|E\phi\|^2_{H^1(\Omega)} - \frac{1}{2}\|\varphi\|^2_{H^1(\Omega_f)}}{\|(\phi,\phi+\varphi)\|_{V_\alpha}\|(\phi,\varphi)\|_{V_\alpha}} \\
&\qquad \geq \frac{\|\phi\|^2_{U_{1+\alpha^{-1}}} + \|\varphi\|^2_{H^1(\Omega_f)} - \frac{1}{2}\|\phi\|^2_{U_{1+\alpha^{-1}}} - \frac{1}{2}\|\varphi\|^2_{H^1(\Omega_f)}}{\|(\phi,\phi+\varphi)\|_{V_\alpha}\|(\phi,\varphi)\|_{V_\alpha}} \\
&\qquad \geq \frac{1}{2\sqrt{3}} \frac{\|(\phi,\varphi)\|^2_{V_\alpha}}{\|(\phi,\varphi)\|_{V_\alpha}\|(\phi,\varphi)\|_{V_\alpha}} \\
&\qquad = \frac{1}{2\sqrt{3}},
\end{aligned}$$

where we have used that $\|(\phi,\phi+\varphi)\|_{V_\alpha} \leq \sqrt{3}\|(\phi,\varphi)\|_{V_\alpha}$, which is a consequence of the triangle inequality and that $2ab \leq (a^2+b^2)$:

$$\begin{aligned}
\|(\phi,\phi+\varphi)\|^2_{V_\alpha} &= \|E\phi\|^2_{H^1(\Omega)} + \frac{1}{\alpha}\|E\phi\|^2_{L^2(\partial\Omega)} + \|\phi+\varphi\|^2_{H^1(\Omega_f)} \\
&\leq \|E\phi\|^2_{H^1(\Omega)} + \frac{1}{\alpha}\|E\phi\|^2_{L^2(\partial\Omega)} \\
&\quad + \|\phi\|^2_{H^1(\Omega_f)} + 2\|\phi\|_{H^1(\Omega_f)}\|\varphi\|_{H^1(\Omega_f)} + \|\varphi\|^2_{H^1(\Omega_f)} \\
&\leq \|(\phi,\varphi)\|^2_{V_\alpha} + \|\phi\|^2_{H^1(\Omega_f)} + \|\phi\|^2_{H^1(\Omega_f)} + \|\varphi\|^2_{H^1(\Omega_f)} \\
&\leq 3\|(\phi,\varphi)\|^2_{V_\alpha}.
\end{aligned}$$

The first inf-sup condition thus holds with $C_2 = \frac{1}{2\sqrt{3}}$. Since (28) is well-posed, the second inf-sup condition immediately follows, with the same constant, from standard Babuška theory. □

We have now verified that the bilinear form $a(\,\cdot\,,\,\cdot\,) : V_\alpha \times V_\alpha \to \mathbb{R}$ is continuous and weakly coercive, with constants that are independent of α. We can then, from the Babuška theory, conclude that

Theorem 6.3. *The operator $\mathcal{A}_\alpha^{2\times 2} : V_\alpha \to V_\alpha'$, defined in (28), is an isomorphism. That is, $\mathcal{A}_\alpha^{2\times 2}$ is bounded and continuously invertible for $\alpha > 0$, in the sense that*

$$\|\mathcal{A}_\alpha^{2\times 2}\|_{\mathcal{L}(V_\alpha, V_\alpha')} \leq C_1 \quad \text{and} \quad \|[\mathcal{A}_\alpha^{2\times 2}]^{-1}\|_{\mathcal{L}(V_\alpha', V_\alpha)} \leq C_2^{-1},$$

where both C_1 and C_2 are independent of $\alpha > 0$.

Proof. The result follows from Lemma 6.1, Lemma 6.2 and standard Babuška theory. □

7 Preconditioners

For a mapping $\mathcal{A}_\alpha$, of the form (16) or (28), from a Hilbert space H onto its dual space H', Krylov subspace methods cannot be applied to solve

$$\mathcal{A}_\alpha x = b.$$

However, assuming that an operator $\mathcal{B}_\alpha : H' \to H$ is available, Krylov subspace methods can be employed to solve

$$\mathcal{B}_\alpha \mathcal{A}_\alpha x = \mathcal{B}_\alpha b, \tag{34}$$

since $\mathcal{B}_\alpha \mathcal{A}_\alpha$ is a mapping from H to H.

In [2, 7, 4] the authors discuss that, to obtain an efficient preconditioner $\mathcal{B}_\alpha$, this mapping should be an isomorphism, with h and α-independent bounds for both $\|\mathcal{B}_\alpha\|$ and $\|\mathcal{B}_\alpha^{-1}\|$. With these ideas in mind, we can propose preconditioners for both the 3×3 and 2×2 block systems analyzed above.

7.1 3×3 system

To construct a suitable preconditioner

$$\mathcal{B}_\alpha^{3\times 3} : F_\alpha' \times U_\alpha' \times U_{\alpha^{-1}}' \to F_\alpha \times U_\alpha \times U_{\alpha^{-1}}$$

for $\mathcal{A}_\alpha^{3\times 3}$, defined in (16), let us recall the scaled norms

$$\begin{aligned} \|f\|_{F_\alpha}^2 &= \alpha \|f\|_{H^1(\Omega_f)}^2, \\ \|u\|_{U_\alpha}^2 &= \alpha \|u\|_{H^1(\Omega)}^2 + \|u\|_{L^2(\partial\Omega)}^2, \\ \|w\|_{U_{\alpha^{-1}}}^2 &= \frac{1}{\alpha} \|w\|_{H^1(\Omega)}^2. \end{aligned}$$

We suggest to use the inverse of the Riesz map of the space

$$W_\alpha = F_\alpha \times U_\alpha \times U_{\alpha^{-1}}$$

as preconditioner, i.e.

$$\mathcal{B}_\alpha^{3\times 3} = \begin{bmatrix} \alpha\tilde{A}_f & 0 & 0 \\ 0 & \alpha A + M_\partial & 0 \\ 0 & 0 & \frac{1}{\alpha}A \end{bmatrix}^{-1} : W_\alpha' \to W_\alpha, \tag{35}$$

see (17)-(20). Clearly, $\mathcal{B}_\alpha^{3\times 3}$ is an isomorphism with α independent bounds for $\|\mathcal{B}_\alpha\|$ and $\|\mathcal{B}_\alpha^{-1}\|$. Moreover, provided that sound discretization techniques are employed, this property will be inherited by the associated discretized operator.

In Section 4 we concluded that both $\|\mathcal{A}_\alpha^{3\times 3}\|$ and $\|[\mathcal{A}_\alpha^{3\times 3}]^{-1}\|$ are bounded independently of $\alpha > 0$. Consequently, also $\|\mathcal{B}_\alpha^{3\times 3}\mathcal{A}_\alpha^{3\times 3}\|$ and $\|[\mathcal{B}_\alpha^{3\times 3}\mathcal{A}_\alpha^{3\times 3}]^{-1}\|$ are well behaved, regardless of the size of $\alpha > 0$. That is, $\mathcal{B}_\alpha^{3\times 3}$ yields a regularization robust preconditioner:

$$\mathcal{B}_\alpha^{3\times 3}\mathcal{A}_\alpha^{3\times 3}x = \mathcal{B}_\alpha^{3\times 3}b. \tag{36}$$

7.2 2×2 system

In the analysis of the operator $\mathcal{A}_\alpha^{2\times 2}$, defined in (28), we used the weighted norm

$$\|u\|_{U_{1+\alpha^{-1}}}^2 = \|Eu\|_{H^1(\Omega)}^2 + \frac{1}{\alpha}\|Eu\|_{L^2(\partial\Omega)}^2,\ u \in H^1(\Omega_f).$$

Thus, a natural preconditioner reads

$$\mathcal{B}_\alpha^{2\times 2} = \begin{bmatrix} \mathsf{A} + \frac{1}{\alpha}\mathsf{M}_\partial & 0 \\ 0 & \mathsf{A}_f \end{bmatrix}^{-1} : V_\alpha' \to V_\alpha, \tag{37}$$

since this is the inverse of the Riesz map of V_α, see (33) and (29)-(32). Sound discretization techniques will, as for the 3×3 system, provide an α-robust preconditioner:

$$\mathcal{B}_\alpha^{2\times 2}\mathcal{A}_\alpha^{2\times 2}x = \mathcal{B}_\alpha^{2\times 2}b. \tag{38}$$

8 Numerical experiments

8.1 The extension operator

In properly weighted Hilbert spaces, the systems (16) and (28) are well-behaved, independently of $\alpha > 0$. To solve these systems numerically, however, we have to represent the subspace $U \subset H^1(\Omega)$, or alternatively, compute the action of the extension operator E, defined in (22). We now address how this can be accomplished.

Let $\{\phi_i\}_{i=1}^N$ be a basis for $V_h \subset H^1(\Omega_f)$, where V_h is a standard scalar FE space. We then get

$$(Eu_h, E\phi_i)_{H^1(\Omega)} = (u_h, \phi_i)_{H^1(\Omega_f)} + (Eu_h, E\phi_i)_{H^1(\Omega_0)}, \tag{39}$$

since the extension leaves the function unchanged throughout Ω_f. Furthermore, with $u_h = \sum_{j=1}^N a_j\phi_j$,

$$(Eu_h, E\phi_i)_{H^1(\Omega_0)} = \sum_{j=1}^N a_j(E\phi_j, E\phi_i)_{H^1(\Omega_0)}.$$

We note from (9)-(11) that $\tilde{\phi}_j = (E\phi_j)|_{\Omega_0}$ is uniquely determined from $\phi_j|_{\partial\Omega_f}$. And,

$$\tilde{\phi}_j = (E\phi_j)|_{\Omega_0} = 0 \text{ if } \phi_j|_{\partial\Omega_f} = 0.$$

Consequently, only the basis functions associated with nodes positioned at the boundary $\partial\Omega_f$ of Ω_f will have non-zero extensions. These extensions are determined by solving (9)-(11). More specifically, one elliptic boundary problem must be solved for each node positioned at $\partial\Omega_f$. This may become CPU demanding if the number of nodes at this interface is large, but the process is easy to parallelize. A more thorough discussion of this issue is present in Section 10.

When the non-zero extensions have been determined, the matrix contributions associated with $(Eu_h, \phi_i)_{H^1(\Omega)}$ can be assembled by computing the two inner-products in (39). And, it is also straightforward to assemble the "boundary" matrix associated with M_∂, see (28)-(29),

$$(Eu_h, E\phi_i)_{L^2(\partial\Omega)} = \sum_{j=1}^N a_j(E\phi_j, E\phi_i)_{L^2(\partial\Omega)}.$$

Remark 8.1. *Krylov subspace solvers typically require that $\mathcal{A}_\alpha^{2\times 2}p$ is computed, for a given (vector) p. Since the extension E is defined in terms of an elliptic PDE, it should be possible to determine $\mathcal{A}_\alpha^{2\times 2}p$ by employing a multigrid scheme, without computing $E\phi_j$ for all indexes associated with nodes at $\partial\Omega_f$. Similarly, it is also likely that the action of the preconditioner $\mathcal{B}_\alpha^{2\times 2}$ can be directly computed with multigrid schemes, see* (37) *and* (29)-(32)*. Hence, one would expect that the step of explicitly determining $\{E\phi_j\}$ can be avoided, provided that proper tailored software is available. We, however, used standard software packages, and it turned out to be difficult to avoid this preprocessing task prior to assembling and solving the KKT system.*

8.2 Numerical setup

To avoid introducing further notation, we will not define new symbols for the matrices and vectors associated with the operators and functions in (16),

(28), (35) and (37). We would like to emphasis that, in this section, all use of these symbols are to the associated discretized versions.

- All simulations were performed using cbc.block; a branch of the FEniCS software [5].
- For all MINRES tests, the preconditioners (35) and (37) were approximated with the Algebraic MultiGrid (AMG) package in PyTrilinos. We used a symmetric Gauss-Seidel smoother, with three smoothing sweeps.
- For the computations of the eigenvalues, presented below, we dumped the matrices to .mat-files and computed the exact preconditioners in Octave.
- In all simulations, we worked on the domains

$$\Omega = (0,1) \times (0,1),$$
$$\Omega_f = (.25,.75) \times (.25,.75).$$

  The observation data d was generated by solving (5) with

$$f(x,y) = 3\cos(\pi x) + y^2, \tag{40}$$

  and setting $d = u|_{\partial\Omega}$, where u denotes the solution of (5).

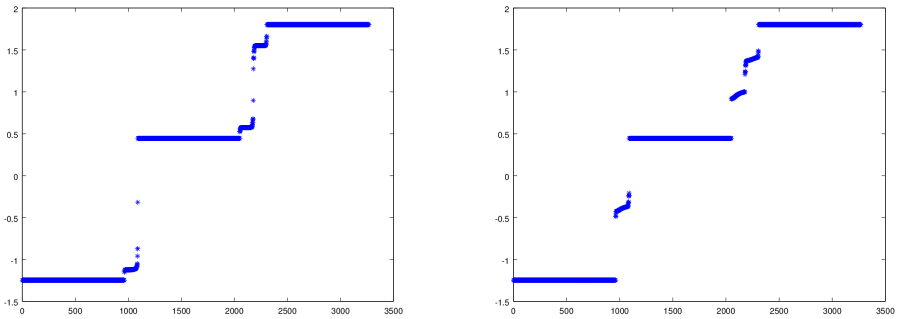
- The MINRES iterations were stopped when

$$\frac{(r_k, \mathcal{B}_\alpha^{p\times p} r_k)}{(r_0, \mathcal{B}_\alpha^{p\times p} r_0)} = \frac{(\mathcal{A}_\alpha^{p\times p} x_k - b, \mathcal{B}^{p\times p}[\mathcal{A}_\alpha^{p\times p} x_k - b])}{(\mathcal{A}_\alpha^{p\times p} x_0 - b, \mathcal{B}^{p\times p}[\mathcal{A}_\alpha^{p\times p} x_0 - b])} \leq \epsilon, \tag{41}$$

  where $p = 2, 3$. In other words, we used a standard relative stopping criterion.

8.3 3×3 system

We first consider the numerical solution of (16) with the preconditioner (35). In Table 1 we can not observe any (systematic) growth in the iteration numbers when the regularization parameter α decreases. The small increase in iteration numbers when the mesh parameter $h \to 0$ is most likely linked to the performance of the AMG. This is supported by Table 2, where we do not observe a significant increase of the condition number $\kappa(\mathcal{B}_\alpha^{3\times 3}\mathcal{A}_\alpha^{3\times 3})$ for $h = 2^{-6}$ compared to $h = 2^{-5}$. That is, for small values of α, the condition number equals 8.701 for $h = 2^{-5}$ and it equals 8.705 for $h = 2^{-6}$. Thus, the discretized preconditioner provides iteration counts, for the MINRES method, which are well behaved with respect to the mesh and regularization parameters.



(a) $\alpha = 1$. (b) $\alpha = 10^{-4}$.

Figure 2: The spectrum of $\mathcal{B}_\alpha^{3\times3}\mathcal{A}_\alpha^{3\times3}$ for two different regularization parameters. These results were computed with a mesh parameter $h = 2^{-6}$.

In Figure 2 the spectrum of $\mathcal{B}_\alpha^{3\times3}\mathcal{A}_\alpha^{3\times3}$ is displayed for two choices of α. Both spectra are clustered, with three large bands of eigenvalues. The remaining eigenvalues seem to become more clustered for smaller values of α than for $\alpha = 1$.

$\alpha\backslash h$	2^{-5}	2^{-6}	2^{-7}	2^{-8}
1	41	45	47	59
10^{-1}	49	54	55	65
10^{-2}	64	70	72	80
10^{-3}	56	66	68	80
10^{-4}	48	52	57	73

Table 1: The number of MINRES iterations required to solve the KKT system (16), i.e. the 3×3 system, with the preconditioner (35). The stopping criterion was $\epsilon = 10^{-8}$.

8.4 2×2 system

To numerically solve the 3×3 block system (16), with the preconditioner (35), we must implement the extension operator E in (22). Thus, it is practically no additional computational effort to reduce the problem to the 2×2 block system (28), with the preconditioner (37). We will now explore how the MINRES algorithm performs on this reduced system.

In Table 3 we observe much smaller iteration numbers than for the 3×3 system, see Table 1. Furthermore, the iteration counts in Table 3 are well behaved with respect to the size of the mesh parameter h, and, if anything, the iteration numbers decreases as α decreases. The latter fact is probably not a generic pattern, but, for this particular model problem, this

α	$\lvert\lambda_1\rvert$	$\lvert\lambda_n\rvert$
1	0.31835	1.8019
10^{-1}	0.22199	1.8019
10^{-2}	0.21079	1.8019
10^{-3}	0.20872	1.8019
10^{-4}	0.20751	1.8019
10^{-5}	0.20713	1.8019
10^{-6}	0.20709	1.8019
10^{-7}	0.20708	1.8019
10^{-8}	0.20708	1.8019

(a) Mesh parameter $h = 2^{-5}$.

α	$\lvert\lambda_1\rvert$	$\lvert\lambda_n\rvert$
1	0.31831	1.8019
10^{-1}	0.22192	1.8019
10^{-2}	0.21078	1.8019
10^{-3}	0.20886	1.8019
10^{-4}	0.20771	1.8019
10^{-5}	0.20710	1.8019
10^{-6}	0.20700	1.8019
10^{-7}	0.20699	1.8019
10^{-8}	0.20699	1.8019

(b) Mesh parameter $h = 2^{-6}$.

Table 2: The smallest and largest eigenvalues, measured in absolute value, of $\mathcal{B}_\alpha^{3\times3}\mathcal{A}_\alpha^{3\times3}$.

observation is supported by the eigenvalues reported in Table 4. That is, the computed condition number $\kappa(\mathcal{B}_\alpha^{2\times2}\mathcal{A}_\alpha^{2\times2})$ is non-increasing as $\alpha \to 0$. The spectrum of $\mathcal{B}_\alpha^{2\times2}\mathcal{A}_\alpha^{2\times2}$ is depicted in Figure 3.

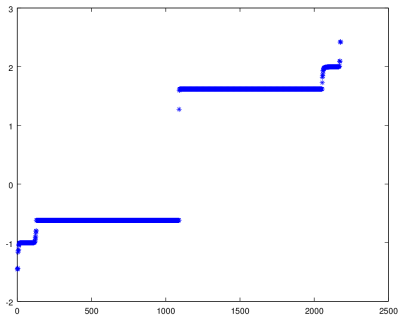
According to standard theory, the MINRES scheme requires $O(\kappa(\mathcal{A}))$ iterations to solve the system $\mathcal{A}x = b$. For the 2×2 block system, with $h = 2^{-6}$ and $\alpha = 10^{-4}$, the condition number is $\kappa(\mathcal{B}_\alpha^{2\times2}\mathcal{A}_\alpha^{2\times2}) = 2.618$, while for the 3×3 block system the condition number is $\kappa(\mathcal{B}_\alpha^{3\times3}\mathcal{A}_\alpha^{3\times3}) = 8.675$, which gives the ratio $8.675/2.618 \approx 3.31$. The ratio between the associated iteration counts is $52/17 \approx 3.06$. Similar results hold for the other choices, reported in our tables, of the mesh parameter and the regularization parameter.

$\alpha\backslash h$	2^{-5}	2^{-6}	2^{-7}	2^{-8}
1	31	31	33	37
10^{-1}	30	30	32	35
10^{-2}	28	27	30	34
10^{-3}	21	20	23	27
10^{-4}	18	17	20	24

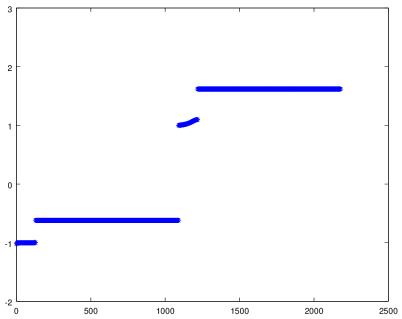
Table 3: The number of MINRES iterations required to solve the KKT system (28), i.e. the 2×2 system, with the preconditioner (37). The stopping criterion was $\epsilon = 10^{-8}$.

8.5 Comparison with standard preconditioners

Recall the original form (4)-(5) of our optimization problem. The KKT system associated with this problem, without invoking the space U of functions



(a) $\alpha = 1$.



(b) $\alpha = 10^{-4}$.

Figure 3: The spectrum of $\mathcal{B}_\alpha^{2\times 2}\mathcal{A}_\alpha^{2\times 2}$ for two different regularization parameters. These results where generated with the mesh parameter $h = 2^{-6}$.

α	$\lvert\lambda_1\rvert$	$\lvert\lambda_n\rvert$
1	0.61803	2.3822
10^{-1}	0.61803	2.2465
10^{-2}	0.61803	1.9123
10^{-3}	0.61803	1.6180
10^{-4}	0.61803	1.6180

(a) Mesh parameter $h = 2^{-5}$.

α	$\lvert\lambda_1\rvert$	$\lvert\lambda_n\rvert$
1	0.61803	2.4246
10^{-1}	0.61803	2.3213
10^{-2}	0.61803	2.0512
10^{-3}	0.61803	1.6180
10^{-4}	0.61803	1.6180

(b) Mesh parameter $h = 2^{-6}$.

Table 4: The smallest and largest eigenvalues, measured in absolute value, of $\mathcal{B}_\alpha^{2\times 2}\mathcal{A}_\alpha^{2\times 2}$.

which are Helmholtz-harmonic on Ω_0, will yield an operator

$$\mathcal{C}_\alpha : H^1(\Omega_f) \times H^1(\Omega) \times H^1(\Omega) \to H^1(\Omega_f)' \times H^1(\Omega)' \times H^1(\Omega)'.$$

Hence, we may simply use the inverse of the Riesz map of $H^1(\Omega_f) \times H^1(\Omega) \times H^1(\Omega)$ to obtain a preconditioned system of the form

$$\mathcal{R}\mathcal{C}_\alpha x = \mathcal{R}b. \tag{42}$$

In this case, no weighting of the involved norms is applied. We will now compare the performance of this methodology with the approaches discussed above. We will not undertake a complete "start-to-finish" comparison, since there are many possibilities, particularly with respect to parallelization, to speed up the computation of the extension operator. Instead, we will perform a pure MINRES test, where we measure the wall-time needed by the different schemes.

Tables 5 and 6 contain the speed-up obtained by solving the 3×3 and 2×2 block systems (36) and (38), respectively, instead of applying the Krylov

subspace solver to (42). As expected, the speed-up increases when $\alpha \to 0$. For example, with $\alpha = 10^{-6}$ and $h < 10^{-5}$, MINRES solves the 2×2 block system more than 48 times faster than the "standard" preconditioned KKT system (42).

$\alpha \backslash h$	2^{-5}	2^{-6}	2^{-7}	2^{-8}
1	1.33	1.71	1.64	1.49
10^{-1}	1.80	2.10	2.22	2.02
10^{-2}	2.00	2.19	2.35	2.42
10^{-3}	3.16	3.88	3.64	3.54
10^{-4}	4.70	6.26	5.17	4.98
10^{-5}	7.22	9.00	8.74	8.58
10^{-6}	9.22	12.5	11.7	10.7

Table 5: Ratio between the wall-time needed by MINRES to solve (42) and (36) (3×3 block system).

$\alpha \backslash h$	2^{-5}	2^{-6}	2^{-7}	2^{-8}
1	2.40	3.63	3.15	3.74
10^{-1}	3.60	5.25	5.24	5.91
10^{-2}	6.50	8.14	7.75	8.86
10^{-3}	12.7	18.6	15.2	16.3
10^{-4}	15.7	23.8	22.7	23.3
10^{-5}	21.7	40.5	39.0	38.1
10^{-6}	27.7	53.0	48.9	48.1

Table 6: Ratio between the wall-time needed by MINRES to solve (42) and (38) (2×2 block system).

Tables 5 and 6 show that the methods introduced in this paper perform favorable compared with the "standard" preconditioning technique. Nevertheless, the comparison is not truly "objective": The stopping criterion depends on the involved operators and on the regularization parameter α, see (41). Therefore, we also performed an alternative comparison, where we added a prior, given by (40), to the minimization problem. More specifically, we replaced the regularization term in the cost functionals (4) and (7) with

$$\frac{1}{2}\alpha \|f - f_{\text{prior}}\|^2_{H^1(\Omega_f)}, \tag{43}$$

where $f_{\text{prior}}(x, y) = 3\cos(\pi x) + y^2$. Hence, f_{prior} is both used to generate the observation data d and as a prior. The control determined by solving the KKT system will therefore be almost equal to f_{prior}. Our objective is to study how fast the approximate controls f_k, generated by the MINRES algorithm, approaches this function.

In Figure 4, the relative difference

$$\frac{\|f_k - f_{\text{prior}}\|_{H^1(\Omega_f)}}{\|f_{\text{prior}}\|_{H^1(\Omega_f)}}$$

is displayed as a function of the number of MINRES iterations k. More specifically, the MINRES method was applied to (42), the 3×3 block system (36) and the 2×2 block system (38). We observe that the relative difference is reduced to 10^{-4} in approximately 220 iterations by the first scheme, while the two latter preconditioning techniques required 35 and 15 iterations, respectively; see figures 4(a), 4(b) and 4(c). In these experiments we used a zero initial guess for the iteration process, $\alpha = 10^{-4}$ and $h = 1/256$.

9 Further analysis

From the analysis presented above, we can conclude that the spectral condition number of the preconditioned operator $\mathcal{B}_\alpha^{2\times 2}\mathcal{A}_\alpha^{2\times 2}$ is bounded independently of α and h. On the other hand, Figure 3 indicates that the spectrum of this operator may possess further nice properties. The purpose of this section is to investigate this issue from an algebraic point of view. Throughout this section we assume that M_∂, A and A_f are FE operators.

A member λ of the point spectrum of $\mathcal{B}_\alpha^{2\times 2}\mathcal{A}_\alpha^{2\times 2}$ must satisfy

$$\underbrace{\begin{bmatrix} -\frac{1}{\alpha}\mathsf{M}_\partial & \mathsf{A} \\ \mathsf{A} & \mathsf{A}_f \end{bmatrix}}_{\mathcal{A}_\alpha^{2\times 2}} \begin{bmatrix} u \\ f \end{bmatrix} = \lambda \underbrace{\begin{bmatrix} \mathsf{A} + \frac{1}{\alpha}\mathsf{M}_\partial & 0 \\ 0 & \mathsf{A}_f \end{bmatrix}}_{(\mathcal{B}_\alpha^{2\times 2})^{-1}} \begin{bmatrix} u \\ f \end{bmatrix},$$

or

$$-\frac{1}{\alpha}\mathsf{M}_\partial u + \mathsf{A}f = \lambda\left(\mathsf{A}u + \frac{1}{\alpha}\mathsf{M}_\partial u\right), \tag{44}$$

$$\mathsf{A}u + \mathsf{A}_f f = \lambda \mathsf{A}_f f. \tag{45}$$

From (45) we find that

$$f = \frac{1}{\lambda - 1}\mathsf{A}_f^{-1}\mathsf{A}u,\ \lambda \neq 1,$$

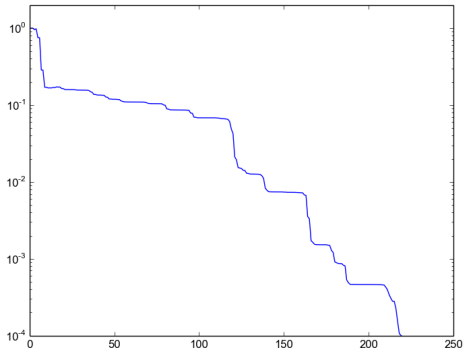
which we may insert into (44) to obtain

$$-\frac{1}{\alpha}(1+\lambda)\mathsf{M}_\partial u + \frac{1}{\lambda - 1}\mathsf{A}\mathsf{A}_f^{-1}\mathsf{A}u - \lambda \mathsf{A}u = 0. \tag{46}$$

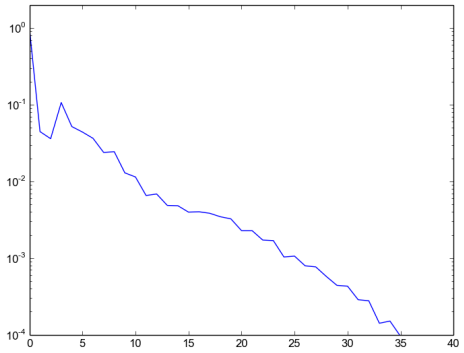
Hence, any eigenvalue must satisfy

$$\frac{1}{\alpha}(1+\lambda)\langle \mathsf{A}^{-1}\mathsf{A}_f\mathsf{A}^{-1}\mathsf{M}_\partial u, u\rangle + \frac{1}{1-\lambda}\langle u, u\rangle + \lambda\langle \mathsf{A}^{-1}\mathsf{A}_f u, u\rangle = 0, \tag{47}$$

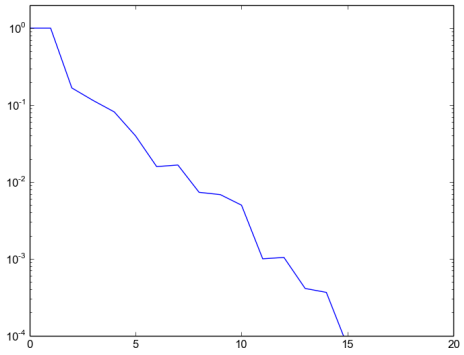
where



(a) "Standard" Riesz preconditioner, see (42).



(b) 3×3 preconditioner, see (35).



(c) 2×2 preconditioner, see (37).

Figure 4: The relative difference $\|f_k - f_{\text{prior}}\|_{H^1(\Omega_f)} / \|f_{\text{prior}}\|_{H^1(\Omega_f)}$ as a function of the number of MINRES iterations k.

- $\mathsf{A}^{-1}\mathsf{A}_f\mathsf{A}^{-1}\mathsf{M}_\partial$ is semi-positive,
- $\mathsf{A}^{-1}\mathsf{A}_f$ is positive.

Theorem 9.1.

a) *Let*

$$\begin{aligned} \underline{\gamma} &= \lambda_{\min}(\mathsf{A}^{-1}\mathsf{A}_f), \\ \overline{\gamma} &= \lambda_{\max}(\mathsf{A}^{-1}\mathsf{A}_f). \end{aligned}$$

Then,

$$\begin{aligned} \mathrm{sp}(\mathcal{B}_\alpha^{2\times 2}\mathcal{A}_\alpha^{2\times 2}) \quad \subset \quad & \left[\min\left\{-1, \frac{1}{2} - \sqrt{\frac{1}{4}+\frac{1}{\underline{\gamma}}}\right\}, \right. \\ & \left. \max\left\{-1, \frac{1}{2} - \sqrt{\frac{1}{4}+\frac{1}{\overline{\gamma}}}\right\}\right] \cup \left(1, \frac{1}{2} + \sqrt{\frac{1}{4}+\frac{1}{\underline{\gamma}}}\right]. \end{aligned}$$

b) *If $\lambda \in \mathrm{sp}(\mathcal{B}_\alpha^{2\times 2}\mathcal{A}_\alpha^{2\times 2})$ is an eigenvalue associated with (u, f), where $u|_{\partial\Omega_f} = 0$, then*

$$\lambda = \frac{1 \pm \sqrt{5}}{2} \approx \begin{cases} 1.618, \\ -0.618. \end{cases}$$

c) *The multiplicity of the eigenvalue $\lambda = -1$ equals the dimension of the null-space of the operator*

$$Q : u \to (u, \phi)_{H^1(\Omega_f)} - (Eu, E\phi)_{H^1(\Omega_0)}, \quad \phi \in V_h,$$

where V_h is a FE space for $H^1(\Omega_f)$.

Remarks

For the model problem associated with Figure 3, $\overline{\gamma} = 1.000$ and $\underline{\gamma} = 0.250$. Hence, in this case, invoking a) yields that

$$\mathrm{sp}(\mathcal{B}_\alpha^{2\times 2}\mathcal{A}_\alpha^{2\times 2}) \subset [-1.562, -0.618] \cup (1, 2.562] \quad \forall \alpha > 0.$$

The eigenvalues $\lambda = -0.618, 1.618$, see b), typically have large multiplicity and appear as "long horizontal line segments" in Figure 3. (These numbers coincide with those derived in [8]. We have not fully investigated the connection between the analysis presented in this paper and [8].)

If $\mathrm{supp}(u) \subset \Omega_f$, then $Eu|_{\Omega_0} = 0$ and, hence, $u \neq 0$ can not belong to the null-space of Q, see c). From this we may conclude that the dimension of the kernel of Q, and thus the multiplicity of $\lambda = -1$, is less or equal to the number of nodes at the boundary $\partial\Omega_f$ of Ω_f.

Proof of Theorem 9.1a).

- Note that, if $u = 0$, then (44) implies that $f = 0$. Hence, there does not exist any eigenfunction (u, f) with $u = 0$. Therefore, in the analysis presented below, we can always assume that $u \neq 0$ in (47).
- Also, for $\lambda = 0$, $(1 + \lambda)$, $\frac{1}{1-\lambda} > 0$ and from (47) we find that 0 can not be an eigenvalue.

Positive eigenvalues

- Since $(1 + \lambda)$, $\frac{1}{1-\lambda}, \lambda > 0$ for $\lambda \in (0, 1)$, (47) yields that the open unit interval $(0, 1)$ contains no eigenvalues.
- For $\lambda = 1$, (45) implies that $u = 0$ and it follows from (44) that also $f = 0$. We conclude that 1 does not belong to the spectrum of $\mathcal{B}_\alpha \mathcal{A}_\alpha$.
- For $\lambda > 1$ we find that, see (47),

$$\begin{aligned} \frac{1}{\alpha}(1+\lambda)\langle \mathsf{A}^{-1}\mathsf{A}_f\mathsf{A}^{-1}\mathsf{M}_\partial u, u\rangle &+ \frac{1}{1-\lambda}\langle u, u\rangle + \lambda\langle \mathsf{A}^{-1}\mathsf{A}_f u, u\rangle \\ &\geq \frac{1}{1-\lambda}\langle u, u\rangle + \lambda\underline{\gamma}\langle u, u\rangle \\ &= \left(\frac{1}{1-\lambda} + \lambda\underline{\gamma}\right)\langle u, u\rangle \\ &> 0 \end{aligned}$$

  if $\lambda > \frac{1}{2} + \sqrt{\frac{1}{4} + \frac{1}{\underline{\gamma}}}$. Thus, there are no eigenvalues larger than $\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{1}{\underline{\gamma}}}$.

We conclude that positive eigenvalues must belong to the interval

$$\left(1, \frac{1}{2} + \sqrt{\frac{1}{4} + \frac{1}{\underline{\gamma}}}\right].$$

Negative eigenvalues

- For $\lambda \in (-1, 0)$ it follows that $(1 + \lambda)$, $\frac{1}{1-\lambda} > 0$ and $\lambda < 0$. Consequently, see (47),

$$\begin{aligned} \frac{1}{\alpha}(1+\lambda)\langle \mathsf{A}^{-1}\mathsf{A}_f\mathsf{A}^{-1}\mathsf{M}_\partial u, u\rangle &+ \frac{1}{1-\lambda}\langle u, u\rangle + \lambda\langle \mathsf{A}^{-1}\mathsf{A}_f u, u\rangle \\ &\geq \frac{1}{1-\lambda}\langle u, u\rangle + \lambda\overline{\gamma}\langle u, u\rangle \\ &= \left(\frac{1}{1-\lambda} + \lambda\overline{\gamma}\right)\langle u, u\rangle \\ &> 0 \end{aligned}$$

  if $\lambda > \frac{1}{2} - \sqrt{\frac{1}{4} + \frac{1}{\overline{\gamma}}}$. Hence, there can not be any eigenvalues in the interval $\left(\max\left\{-1, \frac{1}{2} - \sqrt{\frac{1}{4} + \frac{1}{\overline{\gamma}}}\right\}, 0\right)$.

- For $\lambda < -1$ we find that $1+\lambda$, $\lambda < 0$ and $\frac{1}{1-\lambda} > 0$. Then, (47) yields that

$$\begin{aligned}\frac{1}{\alpha}(1+\lambda)\langle \mathsf{A}^{-1}\mathsf{A}_f\mathsf{A}^{-1}\mathsf{M}_\partial u, u\rangle &+ \frac{1}{1-\lambda}\langle u,u\rangle + \lambda\langle \mathsf{A}^{-1}\mathsf{A}_f u, u\rangle \\ &\leq \frac{1}{1-\lambda}\langle u,u\rangle + \lambda\underline{\gamma}\langle u,u\rangle \\ &= \left(\frac{1}{1-\lambda}+\lambda\underline{\gamma}\right)\langle u,u\rangle \\ &< 0\end{aligned}$$

  if $\lambda < \frac{1}{2} - \sqrt{\frac{1}{4}+\frac{1}{\underline{\gamma}}}$. Therefore, no eigenvalues are less than $\min\left\{-1, \frac{1}{2} - \sqrt{\frac{1}{4}+\frac{1}{\underline{\gamma}}}\right\}$.

We conclude that negative eigenvalues must belong to the interval

$$\left[\min\left\{-1, \frac{1}{2} - \sqrt{\frac{1}{4}+\frac{1}{\underline{\gamma}}}\right\}, \max\left\{-1, \frac{1}{2} - \sqrt{\frac{1}{4}+\frac{1}{\overline{\gamma}}}\right\}\right].$$

This finishes the proof of a). □

Proof of Theorem 9.1b). If $u|_{\partial\Omega_f} = 0$, then the extension operator E generates a zero extension, i.e. $Eu|_{\Omega_0} = 0$. Consequently,

$$\mathsf{M}_\partial u = 0,$$

see (29). Hence, (46) takes the form

$$\frac{1}{\lambda-1}\mathsf{A}\mathsf{A}_f^{-1}\mathsf{A}u - \lambda\mathsf{A}u = 0$$

or

$$\frac{1}{\lambda-1}\mathsf{A}u - \lambda\mathsf{A}_f u = 0. \tag{48}$$

Since $Eu|_{\Omega_0} = 0$ and $Eu|_{\Omega_f} = u$, we find from (31) and (32) that

$$\langle \mathsf{A}u, u\rangle = (Eu, Eu)_{H^1(\Omega)} = (u,u)_{H^1(\Omega_f)} = \langle \mathsf{A}_f u, u\rangle,$$

which, combined with (48), yields that

$$\left(\frac{1}{\lambda-1} - \lambda\right)\langle \mathsf{A}_f u, u\rangle = 0.$$

Recall that A_f is positive, and therefore λ must satisfy

$$\frac{1}{\lambda-1} - \lambda = 0$$

or $\lambda = (1 \pm \sqrt{5})/2$. □

Proof of Theorem 9.1c). Let (u, f) be an eigenvector with eigenvalue $\lambda = -1$. Then, (44) becomes

$$\mathsf{A}f = -\mathsf{A}u$$

or

$$f = -u.$$

Inserting this into (45) yields that

$$\mathsf{A}u - \mathsf{A}_f u = \mathsf{A}_f u,$$

which we may express on the form

$$\mathsf{A}u - 2\mathsf{A}_f u = 0.$$

This implies that, see (31) and (32),

$$(Eu, E\phi)_{H^1(\Omega)} - 2(u, \phi)_{H^1(\Omega_f)} = 0 \quad \forall \phi \in V_h.$$

Since $Eu|_{\Omega_f} = u$ and $\Omega_0 = \Omega \setminus \Omega_f$, it follows that

$$(Eu, E\phi)_{H^1(\Omega_0)} - (u, \phi)_{H^1(\Omega_f)} = 0 \quad \forall \phi \in V_h. \tag{49}$$

Evidently, if $u \neq 0$ solves (49) and $f = -u$, then (u, f) will satisfy (44)-(45) with $\lambda = -1$. This completes the argument. □

10 Summary, discussion and conclusions

We have introduced a robust preconditioner for a PDE-constrained optimization problem with local control and with boundary observations only. This extends previous results, which have mainly focused on optimization tasks for globally defined controls and global observations.

The state equation of our model problem is elliptic, and the robust preconditioning strategy is derived by employing the "natural" Hilbert space for this equation. More specifically, the solution of the state equation is Helmholtz-harmonic on the complement of the support $\Omega_f \subset\subset \Omega$ of the control. Consequently, the Sobolev norm of the functions belonging to this Hilbert space is equivalent to the Sobolev norm associated with Ω_f. Based on this observation, we can define a preconditioner which is robust with respect to the size of the regularization parameter $\alpha > 0$. Furthermore, this approach enables us to significantly reduce the size of the KKT system: All unknowns are only defined on the support Ω_f of the control, and the Lagrange multiplier can be removed from the problem, which yields a 2×2 block-system, instead of a 3×3 system. For our model problem, employing rather fine meshes and with a relatively small regularization parameter, we obtained a reduction of the computing time by a factor in the range $[20, 50]$ - only recording the CPU-time needed to solve the KKT systems.

Prior to solving the 2×2 block system, a Helmholtz-harmonic extension of each FEM basis function, associated with nodes at $\partial\Omega_f$, must be computed. This is, of course, a negative aspect of our methodology. But this process can be fully parallelized with optimal speed-up. Also, if purely serial computations are employed, one will typically have very good initial guesses for an iterative scheme, e.g. CG, for computing these extensions: The Helmholtz-harmonic extensions of neighboring FEM basis functions can be used as initial guesses. Moreover, the benefits of our scheme increases as the size of the support Ω_f of the control decreases, because fewer extensions must be determined and the reduction of the number of unknowns increases.

From an inverse problem perspective, it is certainly not sufficient to solve the KKT system once with one particular choice of the size of the regularization parameter α. On the contrary, since the noise level of the data typically is unknown, a series of KKT systems, with varying degree of regularization, must be solved in order to determine a close-to-optimal value for α, see e.g. [3]. In such a process, it is only necessary to compute the above-mentioned Helmholtz-harmonic extensions once, and the speed-up obtained by solving the 2×2 KKT system, using our α robust preconditioner, will be large. A similar beneficial situation will occur if it is desirable to solve the PDE-constrained optimization problem with a number of different observation data sets, i.e. solving many KKT systems with different data d. This will be the case for a number of inverse problems arising in the engineering sciences. For example, for the inverse problem of electrocardiography (ECG).

By employing tailored multigrid schemes, we also expect that it might be possible to avoid the preprocessing step of explicitly computing the Helmholtz-harmonic extensions. This turned out to be very difficult to achieve by using standard multigrid software packages, and must be regarded as an open problem.

Our investigation only concerns elliptic control problems, and we assume that the unknown control belongs to H^1. If one insists on searching for a L^2-control, which would allow discontinuities, we do not know how to construct a robust preconditioner for the associated KKT system, but one could try to generalize the approach presented in [6]. It is also an open question how to obtain similar results for parabolic state equations or other relevant PDE models.

References

[1] I. Babuška. Error-bounds for finite element method. *Numerische Mathematik*, 16(19):322–333, 1971.

[2] A. Günnel, R. Herzog, and E. Sachs. A note on preconditioners and scalar products in Krylov subspace methods for self-adjoint problems in

Hilbert space. *Electronic Transactions on Numerical Analysis*, 41:13–20, 2014.

[3] P. C. Hansen. *Discrete Inverse Problems: Insight and Algorithms*. SIAM, 2010.

[4] R. Herzog and E. Sachs. Superlinear convergence of Krylov subspace methods for self-adjoint problems in Hilbert space. Technische Universität Chemnitz, 2014.

[5] K. A. Mardal and J. B. Haga. Block preconditioning of systems of PDEs. In A. Logg, K. A. Mardal, and G. Wells, editors, *Automated Solution of Differential Equations by the Finite Element Method*, pages 643–654. Springer, 2012.

[6] K. A. Mardal, B. F. Nielsen, and Nordaas. M. Robust preconditioners for PDE-constrained optimization with limited observations. *Submitted*, 2014.

[7] K. A. Mardal and R. Winther. Preconditioning discretizations of systems of partial differential equations. *Numerical Linear Algebra with Applications*, 18(1):1–40, 2011.

[8] M. F Murphy, G. H. Golub, and A. J. Wathen. A note on preconditioning for indefinite linear systems. *SIAM Journal on Scientific Computing*, 21(6):1969–1972, 2000.

[9] B. F. Nielsen and K. A. Mardal. Efficient preconditioners for optimality systems arising in connection with inverse problems. *SIAM Journal on Control and Optimization*, 48(8):5143–5177, 2010.

[10] J. W. Pearson, M. Stoll, and A. J. Wathen. Regularization-robust preconditioners for time-dependent PDE-constrained optimization problems. *SIAM Journal on Matrix Analysis and Applications*, 33:1126–1152, 2012.

[11] J. W. Pearson and A. J. Wathen. A new approximation of the Schur complement in preconditioners for PDE-constrained optimization. *Numerical Linear Algebra with Applications*, 19:816–829, 2012.

[12] J. Schöberl and W. Zulehner. Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems. *SIAM Journal on Matrix Analysis and Applications*, 29(3):752–773, 2007.

[13] W. Zulehner. Nonstandard norms and robust estimates for saddle point problems. *SIAM Journal on Matrix Analysis and Applications*, 32:536–560, 2011.

5

Paper IV - Analysis of the Diffuse Domain Method for second order elliptic boundary value problems

This paper is submitted for publication.

Analysis of the Diffuse Domain Method for second order elliptic boundary value problems

Martin Burger[*], Ole Løseth Elvetun[†] and Matthias Schlottbom[‡]

December 17, 2014

Abstract

The diffuse domain method for partial differential equations on complicated geometries recently received strong attention in particular from practitioners, but many fundamental issues in the analysis are still widely open. In this paper we study the diffuse domain method for approximating second order elliptic boundary value problems posed on bounded domains, and show convergence and rates of the approximations generated by the diffuse domain method to the solution of the original second order problem when complemented by Robin, Dirichlet or Neumann conditions.

The main idea of the diffuse domain method is to relax these boundary conditions by introducing a family of phase-field functions such that the variational integrals of the original problem are replaced by a weighted average of integrals of perturbed domains. From an functional analytic point of view, the phase-field functions naturally lead to weighted Sobolev spaces for which we present trace and embedding results as well as various type of Poincaré inequalities with constants independent of the domain perturbations. Our convergence analysis is carried out in such spaces as well, but allows to draw conclusions also about unweighted norms applied to restrictions on the original domain. Our convergence results are supported by numerical examples.

Keywords: Diffuse domain method, weighted Sobolev spaces, domain perturbations, elliptic boundary value problems

AMS subject classifications: 35J20, 35J70, 46E35, 65N85

[*]Institute for Computational and Applied Mathematics, University of Münster, Einsteinstr. 62, 48149 Münster, Germany; Cells in Motion Cluster of Excellence, University of Münster. E-mail: burger@uni-muenster.de

[†]Dept. of Mathematical Sciences and Technology, Norwegian University of Life Sciences. E-mail: ole.elvetun@nmbu.no

[‡]Institute for Computational and Applied Mathematics, University of Münster, Einsteinstr. 62, 48149 Münster, Germany. E-mail: schlottbom@uni-muenster.de

1 Introduction

This paper considers the approximation properties of the diffuse domain method (also called diffuse interface method, cf. [24, 33]) when applied to linear second order elliptic equations of the form

$$-\mathrm{div}(A\nabla u) + cu = f \quad \text{in } D \tag{1}$$

complemented by suitable boundary conditions on a sufficiently smooth domain $D \subset \mathbb{R}^n$. We focus on Neumann, Robin and Dirichlet boundary conditions, i.e. either

$$n \cdot A\nabla u + bu = g \quad \text{on } \partial D \qquad \text{or} \tag{2}$$

$$u = g \quad \text{on } \partial D. \tag{3}$$

For solving equations of the above type we will employ variational methods. Roughly speaking, for instance (1)–(2) is reformulated as follows: Find u such that for all suitable test functions v

$$\int_D A\nabla u \cdot \nabla v + cuv \,\mathrm{d}x + \int_{\partial D} buv \,\mathrm{d}\sigma = \int_D fv \,\mathrm{d}x + \int_{\partial D} gv \,\mathrm{d}\sigma. \tag{4}$$

In many applications the domain D, the boundary ∂D (or equally well some interface inside the domain) is not known exactly or its geometry is complicated making a proper approximation of the integrals difficult or expensive [2, 13, 16, 29, 32, 35]. In addition to the methods used in the aforementioned references, let us point to further literature dealing with methods to handle these type of difficulties; for instance the immersed boundary method [31], the immersed interface method [23], the fictitious domain method [15], the unfitted finite element method [5], the finite cell method [30], unfitted discontinuous Galerkin methods [6], composite finite elements [19, 25]; let us also refer to these papers for further links to literature and applications. In this work we will focus on the diffuse domain method, see for instance [24]. The diffuse domain method relies on the fact that the domain D can be described by its oriented distance function $d_D(x) = \mathrm{dist}(x, D) - \mathrm{dist}(x, \mathbb{R}^n \setminus D)$, $x \in \mathbb{R}^n$. As one can easily see, we have $D = \{d_D < 0\}$. In order to relax the sharp interface condition $d_D < 0$, let us introduce $\varphi^\varepsilon = S(-d_D/\varepsilon)$ for $\varepsilon > 0$ small and S being a sigmoidal function, i.e. non-decreasing with $S(t) = t/|t|$ for $|t| \geq 1$. As ε tends to zero, $S(\cdot/\varepsilon)$ converges to the sign function, and hence the phase-field function $\omega^\varepsilon = (1+\varphi^\varepsilon)/2$ formally converges to the indicator function χ_D of D. The key idea to approximate the integrals in (4) is a weighted averaging of the integrals over $\{d_D < t\}$ instead of integrating over the original domain $\{d_D < 0\}$ only (and similar for boundary integrals).

Since $\frac{1}{2\varepsilon}S'(\frac{\cdot}{\varepsilon})$ approximates a concentrated distribution at zero, we expect

$$\begin{aligned}\int_D h(x)\,\mathrm{d}x &= \int_{\{d_D<0\}} h(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} \frac{1}{2\varepsilon}S'(-\frac{t}{\varepsilon}) \int_{\{d_D<0\}} h(x)\,\mathrm{d}x\,\mathrm{d}t \\ &\approx \int_{-\infty}^{\infty} \frac{1}{2\varepsilon}S'(-\frac{t}{\varepsilon}) \int_{\{d_D<t\}} h(x)\,\mathrm{d}x\,\mathrm{d}t \\ &= \frac{1}{2}\int_{-1}^{1}\int_{\{\varphi^\varepsilon>s\}} h(x)\,\mathrm{d}x\,\mathrm{d}s,\end{aligned}$$

where we have used the substitution $s = S(-\frac{t}{\varepsilon})$ in the last step. Now the layer cake-representation can further be used for given integrable h to rewrite

$$\int_{-1}^{1}\int_{\{\varphi^\varepsilon>s\}} h(x)\,\mathrm{d}x\,\mathrm{d}t = \int_\Omega \int_{-1}^{\varphi^\varepsilon(x)} \mathrm{d}s\, h(x)\,\mathrm{d}x = \int_\Omega (1+\varphi^\varepsilon(x))h(x)\,\mathrm{d}x.$$

By an analogous computation we obtain for the boundary integral

$$\int_{\partial D} h(x)\,\mathrm{d}\sigma(x) \approx \frac{1}{2}\int_{-1}^{1}\int_{\partial\{\varphi^\varepsilon>s\}} h(x)\,\mathrm{d}\sigma(x)\,\mathrm{d}s,$$

which can be simplified via the co-area formula to

$$\int_{-1}^{1}\int_{\partial\{\varphi^\varepsilon>s\}} h(x)\,\mathrm{d}\sigma(x)\,\mathrm{d}t = \int_\Omega h(x)|\nabla\varphi^\varepsilon(x)|\,\mathrm{d}x.$$

Here, $\Omega \supset \overline{D}$ is a domain with "simple" geometry, i.e. a geometry which can be easily approximated. Based on this motivation we shall define the following diffuse volume and surface integrals

$$\int_D h(x)\,\mathrm{d}x \approx \int_\Omega h(x)\omega^\varepsilon(x)\,\mathrm{d}x \quad\text{and}\quad \int_{\partial D} h(x)\,\mathrm{d}\sigma(x) \approx \int_\Omega h(x)|\nabla\omega^\varepsilon(x)|\,\mathrm{d}x. \tag{5}$$

Using this approximation in (4) leads us to the following variational problem: Find u^ε such that for all suitable test functions v

$$\begin{aligned}\int_\Omega \left(A\nabla u^\varepsilon\cdot\nabla v + cu^\varepsilon v\right)\omega^\varepsilon\,\mathrm{d}x + \int_\Omega bu^\varepsilon v|\nabla\omega^\varepsilon|\,\mathrm{d}x \\ = \int_\Omega fv\omega^\varepsilon\,\mathrm{d}x + \int_\Omega gv|\nabla\omega^\varepsilon|\,\mathrm{d}x.\end{aligned} \tag{6}$$

Under the usual assumptions on A, b and c, the bilinear form on the left-hand side of (6) is well-defined on the weighted Sobolev space $W^{1,2}(\Omega;\omega^\varepsilon)$, which is the closure of smooth functions $u:\Omega\to\mathbb{R}$ with finite (semi-) norm

$$\|u\|^2_{W^{1,2}(\Omega;\omega^\varepsilon)} = \int_\Omega \left(|\nabla u|^2 + |u|^2\right)\omega^\varepsilon\,\mathrm{d}x.$$

The main point of the present manuscript is to estimate the error between the solution u of (4) and the solution u^ε of (6). Our key results are the following two theorems, the first treating the low regularity case and the second giving optimal rates under full regularity:

Theorem 1.1. *Let ∂D be of class $C^{1,1}$ and let $0 \leq c \in L^\infty(\Omega)$, $0 \leq b \in W^{1,\infty}(\Omega)$ and $A \in L^\infty(\Omega)$ such that $\kappa \leq A(x) \leq \kappa^{-1}$ for all $x \in \Omega$ and some constant $\kappa > 0$. Moreover, assume that $f \in L^2(\Omega;\omega^\varepsilon)$ and $g \in W^{1,2}(\Omega;\omega^\varepsilon)$. Furthermore, let $u \in W^{1,2}(D)$ be a solution to (4) and $u^\varepsilon \in W^{1,2}(\Omega;\omega^\varepsilon)$ be a solution to (6). Then there exists $p > 2$ and a constant $C > 0$ independent of ε such that*

$$\|u - u^\varepsilon\|_{W^{1,2}(\Omega;\omega^\varepsilon)} \leq C\varepsilon^{\frac{1}{2}-\frac{1}{p}}.$$

Theorem 1.2. *In addition to the assumptions of Theorem 1.1 let ∂D be of class C^∞, and let $f, g, A, b, c \in C^\infty(\overline{\Omega})$. Then there exists a constant C independent of ε such that*

$$\|u - u^\varepsilon\|_{W^{1,2}(\Omega;\omega^\varepsilon)} \leq C\varepsilon^{3/2}.$$

Let us mention that the case $b = 0$ is allowed here and corresponds to Neumann boundary conditions. The index p in Theorem 1.1 is related to L^p regularity of ∇u, see [18, 26] and Section 6.1 below. In order to prove the theorems, we need a few technical ingredients. As is obvious from the above discussion, we have to deal with a family of weighted Sobolev spaces parametrized by ε. For certain choices of S, we observe that the weight ω^ε is proportional to a power of a distance function near ∂D, see (12) below. For this type of weights and fixed ε many results have been established in literature; see for instance [20, 21, 27, 28] and more generally [36]. Since we are dealing with a family of spaces corresponding to a family of weights ω^ε, we are particularly interested in the behavior of the weighted spaces when ε changes. We will present trace theorems, embedding theorems and Poincaré inequalities with constants independent of ε, which turn out to be indispensable tools for the analysis of (6), and which we think might be of interest in their own, see Section 4. In order to prove these statements, we have to revise and adapt the classical proofs of [28] and combine them with arguments recently used in the context of shape optimization; see [3, 8, 11] for such arguments applied to unweighted Sobolev spaces.
A further necessary ingredient to obtain the error estimates of Theorem 1.1 and Theorem 1.2 are rigorous error estimates for the approximations (5) in terms of ε and the regularity of the integrands. A consequence of our results is that for $h \in L^p(\Omega;\omega^\varepsilon)$

$$\int_D h \,\mathrm{d}x - \int_\Omega h\omega^\varepsilon \,\mathrm{d}x = O(\varepsilon^{1-1/p}), \qquad \text{as } \varepsilon \to 0,$$

see Theorem 5.1. Assuming h and ∇h in $L^p(\Omega;\omega^\varepsilon)$, we can exploit the special averaging procedure in the derivation of the diffuse integrals in a crucial way to obtain the improved estimate

$$\int_D h \,\mathrm{d}x - \int_\Omega h\omega^\varepsilon \,\mathrm{d}x = O(\varepsilon^{2-1/p}), \qquad \text{as } \varepsilon \to 0,$$

cf. Theorem 5.2. Concerning Robin boundary values, let us mention recent formal result obtained by asymptotic analysis stating an L^2-convergence rate for the error $u - u^\varepsilon$ of $O(\varepsilon^2)$ [22]. Using a problem adapted norm we also obtain a rate $O(\varepsilon^2)$. For two-dimensional problems and under reasonable assumptions on this problem adapted norm, which we however cannot verify for our problem, we also arrive at a L^2-convergence rate $O(\varepsilon^2)$. The latter is well confirmed by numerical results.
Concerning Dirichlet boundary values, the corresponding convergence rate only yields the half order compared to the Robin boundary values. Hence in the setting of Theorem 1.1, but with (3) instead of (2), we obtain

$$\|u - u^\varepsilon\|_{W^{1,2}(D)} = O(\varepsilon^{\frac{1}{4}-\frac{1}{2p}}), \qquad \text{as } \varepsilon \to 0,$$

and accordingly, in the setting of Theorem 1.2 with (3) in place of (2), we obtain

$$\|u - u^\varepsilon\|_{W^{1,2}(D)} = O(\varepsilon^{\frac{3}{4}}), \qquad \text{as } \varepsilon \to 0.$$

In the best case, using the problem adapted norm, we can show a rate $O(\varepsilon)$. This complies with recent results in literature, which were obtained for one-dimensional problems or numerically [14, 34].
The outline of the manuscript is as follows. In Section 2 we discuss the geometry of D and certain perturbations of it. Section 3 introduces weighted Lebesgue and Sobolev spaces together with some basic properties. A more detailed study of weighted Sobolev spaces is presented in Section 4. Here we present a trace and an embedding theorem for weighted Sobolev spaces and we show that the corresponding estimates are stable with respect to ε. Moreover, we prove Poincaré and Poincaré-Friedrichs inequalities for these spaces again with constants independent of ε. Approximation properties of the diffuse integrals are presented in Section 5. The volume integrals are investigated in Section 5.1, and corresponding results for the diffuse boundary integral are subsequently shown in Section 5.2. Section 6 deals with the approximation of elliptic equations by the diffuse domain method for Robin, Dirichlet and Neumann type boundary values, and proofs of Theorem 1.1 and Theorem 1.2 are given. Our results are supported by numerical results which are presented in Section 7. We conclude in Section 8 and discuss briefly some open questions.

2 Some geometric preliminaries

2.1 Domain

Throughout the manuscript we assume that $D \subset \mathbb{R}^n$ is a domain with $C^{1,1}$ boundary. Associated to D we define its oriented distance function d_D by

$$d_D(x) = \operatorname{dist}(x, D) - \operatorname{dist}(x, \mathbb{R}^n \setminus D) \quad \text{for } x \in \mathbb{R}^n.$$

Since ∂D is of class $C^{1,1}$ we have that d_D is $C^{1,1}$ in a neighborhood of ∂D [11]. For $t \in \mathbb{R}$ let us define the sublevel sets of d_D as follows

$$D_t = \{x \in \mathbb{R}^n : \ d_D(x) < t\}.$$

We clearly have the inclusions $D_{t_1} \subset D_0 = D \subset D_{t_2}$ for all $t_1 < 0 < t_2$. Moreover, we fix a domain $\Omega \subset \mathbb{R}^n$ such that $\overline{D} \subset \Omega$. In applications Ω is a domain with a "simple" geometry, for instance a ball or a bounding box.

2.2 The tubular neighborhood Γ_ε

Let us define the ε-tubular neighborhood of ∂D by

$$\Gamma_\varepsilon = D_\varepsilon \setminus \overline{D_{-\varepsilon}}. \tag{7}$$

Due to $C^{1,1}$ regularity of ∂D, the projection of $z \in \Gamma_\varepsilon$ onto ∂D is unique for ε sufficiently small, i.e., for each $z \in \Gamma_\varepsilon$ there exists a unique $x \in \partial D$ such that $z = x + d_D(z)n(x)$ [11, Chapter 7, Theorem 3.1, Theorem 8.4]. Here, $n(x)$, $x \in \partial D$, denotes the outward unit normal field which is related to the oriented distance function via the formula $n(x) = \nabla d_D(x)$. This shows

$$\Gamma_\varepsilon = \{z \in \Omega : \exists x \in \partial D, |t| < \varepsilon, z = x + tn(x)\}.$$

In the whole manuscript we fix ε_0 so small such that the just described projection $\Gamma_{2\varepsilon_0} \to \partial D$ is single-valued. Thus, for each $\varepsilon \leq \varepsilon_0$, $D_\varepsilon = \Gamma_\varepsilon \cup D$, and for every $x \in \Gamma_\varepsilon$, there holds $\operatorname{dist}(x, \partial D) \leq \varepsilon$ and for every $x \in \Omega \setminus \Gamma_\varepsilon$, there holds $\operatorname{dist}(x, \partial D) \geq \varepsilon$. Moreover, for some constant $C > 0$ independent of ε

$$|\Gamma_\varepsilon| \leq C\mathcal{H}^{n-1}(\partial D)\varepsilon. \tag{8}$$

Here, $|\Gamma_\varepsilon| = \mathcal{L}^n(\Gamma_\varepsilon)$ is the n-dimensional Lebesgue-measure of Γ_ε and $\mathcal{H}^{n-1}(\partial D)$ is the $n-1$-dimensional Hausdorff-measure of ∂D.

2.3 Transformations of the geometry

Let $t \in (-\varepsilon_0, \varepsilon_0)$. Let us first consider transformations of boundaries $\partial D \to \partial D_t$. To do so, we introduce a family of transformations $\Phi_t : \partial D \to \partial D_t$

defined by $\Phi_t(x) = x+tn(x)$. The Jacobian satisfies $D\Phi_t(x) = I+tD^2d_D(x)$, with I being the identity matrix on $\mathbb{R}^{n\times n}$ and D^2d_D being the Hessian of d_D, and thus,

$$|\det D\Phi_t(x) - (1 + t\Delta d_D(x))| \leq C\|D^2d_D\|_{L^\infty(\partial D)}t^2 \quad \text{for } |t| \leq \varepsilon \leq \varepsilon_0. \tag{9}$$

Decreasing ε_0 if necessary, there holds $\frac{1}{2} \leq \det D\Phi_t(x) \leq 2$ and

$$|\det D\Phi_t(x) - \det D\Phi_{-t}(x)| \leq C|t|\|D^2d_D\|_{L^\infty(\partial D)}. \tag{10}$$

Denoting by n_t the unit outer normal vector field of ∂D_t, we see that $n_t(\Phi_t(x)) = n(x)$ for all $x \in \partial D$ by the choice of the tubular neighborhood Γ_{ε_0}. Thus, here and in the following, we will just write $n(x)$ for the unit outer normal at some $x \in \partial D_t$. In particular, Φ_t can be extended to the whole of Γ_{ε_0}, and for this extension we have that

$$\begin{aligned}\Phi_t(\Phi_s(x)) &= \Phi_s(x) + tn(\Phi_s(x)) = x + (s+t)n(x)\\ &= \Phi_{s+t}(x), \quad s,t \in (-\varepsilon_0,\varepsilon_0),\end{aligned}$$

particularly, $\Phi_t(\Phi_{-t}(x)) = x$. For $h \in L^1(\Gamma_\varepsilon)$, $-\varepsilon < a < b < \varepsilon$ and $s \in (a,b)$ we then have

$$\int_{\{a<d_D<b\}} h(x)\,\mathrm{d}x = \int_a^b \int_{\partial D_s} h(x + (t-s)n(x))|\det D\Phi_{t-s}(x)|\,\mathrm{d}\sigma_s(x)\,\mathrm{d}t. \tag{11}$$

Note that $x - sn(x) \in \partial D$ for $x \in \partial D_s$, and hence, $x + (t-s)n(x) \in \partial D_t$ for $x \in \partial D_s$. Here, $\sigma_s = \mathcal{H}^{n-1}\llcorner\partial D_s$ is the surface element of ∂D_s, i.e.

$$\sigma_s(\widetilde{\Omega}) = \mathcal{H}^{n-1}(\widetilde{\Omega} \cap \partial D_s) \quad \text{for } \widetilde{\Omega} \subset \mathbb{R}^n.$$

For the volume transformation $D \to D_t$, we define $\psi_t : [0,\infty) \to \mathbb{R}$ by $\psi_t(s) = 0$ for $s \geq \varepsilon_0$, and $\psi_t(s) = \frac{t}{\varepsilon_0^2}s^2 - \frac{2t}{\varepsilon_0}s + t$ for $s < \varepsilon_0$. Then $\psi_t \in C^1([0,\infty))$, and $\|\psi_t\|_{C^1([0,\infty))} \to 0$ as $t \to 0$. Moreover, ψ_t maps $[0,\varepsilon_0]$ one-to-one onto $[0,t]$. We then define the diffeomorphism

$$\Psi_t : D \to D_t, \quad \Psi_t(x) = \begin{cases} x + \psi_t(|d_D(x)|)n(x), & x \in D \cap \Gamma_{\varepsilon_0},\\ x, & x \in D \setminus \Gamma_{\varepsilon_0}.\end{cases}$$

Since $\nabla d_D(x) = n(x)$, the Jacobian of Ψ_t is given by

$$D\Psi_t(x) = I - \psi_t'(|d_D(x)|)n(x) \otimes n(x) + \psi_t(|d_D(x)|)D^2d_D(x),$$

and $\sup_{x\in D_{\varepsilon_0}} \|D\Psi_t(x) - I\| \to 0$ as $t \to 0$ by construction of ψ_t. We note that $\Psi_{t|\partial D} = \Phi_t$.

3 Weighted Sobolev spaces

In order to construct the weighted spaces mentioned in the introduction let us begin with defining another level set function for the domain D resembling a sign function smoothed in Γ_ε, namely

$$\varphi^\varepsilon(x) = S\left(\frac{-d_D(x)}{\varepsilon}\right),$$

where the function S is a regularization of the sign function. Hence, $\varphi^\varepsilon(x) > 0$ if and only if $x \in D$. To be precise, we assume that S verifies the following assumptions.

(S1) $S : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous, $S(t) = t/|t|$ for $|t| \geq 1$, and $S'(t) > 0$ for $|t| < 1$. Moreover, $S(t) = -S(-t)$ for all $t \in \mathbb{R}$.

(S2) There exist $\zeta_1, \zeta_2 > 0$ and $\alpha > 0$ such that for all $t \in (0, 2)$

$$\zeta_1 t^\alpha \leq (1 + S(t-1))/2 \leq \zeta_2 t^\alpha.$$

(S3) $S'(t) \leq S'(s)$ for all $0 \leq s \leq t < 1$.

(S3) asserts concavity of S on $(0, 1)$ and this assumption is only needed in Theorem 5.6 below. Assumption (S2) ensures that that the phase-field function ω^ε is proportional to $\operatorname{dist}(\cdot, \partial D_\varepsilon)^\alpha$ on Γ_ε, where ω^ε is defined as a regularization of the indicator function χ_D of D as follows

$$\omega^\varepsilon(x) = \frac{1}{2}\left(1 + \varphi^\varepsilon(x)\right).$$

Obviously, we have that $D_\varepsilon = \{x \in \Omega : \omega^\varepsilon(x) > 0\}$, and $\omega^\varepsilon(x) = 1$ for $x \in D$ with $\operatorname{dist}(x, \partial D) > \varepsilon$. Let us clarify (S2) in the following. We observe that $\operatorname{dist}(x, \partial D_\varepsilon) = \varepsilon - d_D(x)$ for $x \in \Gamma_\varepsilon$. Thus, by (S2) with $t = \operatorname{dist}(x, \partial D_\varepsilon)/\varepsilon$

$$\zeta_1 \left(\frac{\operatorname{dist}(x, \partial D_\varepsilon)}{\varepsilon}\right)^\alpha \leq \omega^\varepsilon(x) \leq \zeta_2 \left(\frac{\operatorname{dist}(x, \partial D_\varepsilon)}{\varepsilon}\right)^\alpha \tag{12}$$

Before proceeding, let us give a few examples which may serve as prototypes for S.

Example 3.1. *(i) Let $S(t) = t$ for $|t| < 1$. Obviously (S1) and (S3) are satisfied. Moreover, $1 + S(t-1) = t$, and we can choose $\alpha = 1$, and $\zeta_1 = \zeta_2 = 1/2$ in (S2).*
(ii) Let $S(t) = (3t - t^3)/2$. Thus, $S \in C^1(\mathbb{R})$ and $S'(t) = 3(1 - t^2)/2 > 0$ for $|t| < 1$. Since $S''(t) = -3t < 0$ for $t > 0$, (S3) is satisfied. Moreover, $1 + S(t-1) = (3t^2 - t^3)/2$, and (S2) is satisfied for $\alpha = 2$ and $\zeta_1 = 1/4$ and $\zeta_2 = 3/4$.
(iii) Let $S(t) = 15t/8 - 5t^3/4 + 3t^5/8$. Thus, $S \in C^2(\mathbb{R})$ and $S'(t) = 15/8 - 15t^2/4 + 15t^4/8 > 0$ for $|t| < 1$. Since $S''(t) = -15t/2 + 15t^3/2 < 0$ for $t > 0$, (S3) is satisfied. Moreover, $1 + S(t-1) = 5t^3/2 - 15t^4/8 + 3t^5/8$, and (S2) is satisfied for $\alpha = 3$ and $\zeta_1 = 1/8$ and $\zeta_2 = 5/2$.

Proceeding with the construction of weighted Lebesgue spaces, let us introduce the measure

$$\omega^\varepsilon(\widetilde{\Omega}) = \int_{\widetilde{\Omega}} \omega^\varepsilon(x)\,\mathrm{d}x, \quad \widetilde{\Omega} \subset \mathbb{R}^n \text{ measurable},$$

which is absolutely continuous with respect to the Lebesgue measure. Associated to ω^ε let us further introduce for $1 \le p < \infty$ the weighted L^p-spaces

$$L^p(D_\varepsilon;\omega^\varepsilon) = \{v : D_\varepsilon \to \mathbb{R} : \ |v|^p\omega^\varepsilon \in L^1(D_\varepsilon)\}$$

with norm

$$\|v\|^p_{L^p(D_\varepsilon;\omega^\varepsilon)} = \int_{D_\varepsilon} |v|^p\,\mathrm{d}\omega^\varepsilon.$$

For $p = \infty$, we set $L^\infty(D_\varepsilon;\omega^\varepsilon) = L^\infty(D_\varepsilon)$, i.e. $L^\infty(D_\varepsilon;\omega^\varepsilon)$ is the class of Lebesgue-measurable functions being essentially bounded. In the following we will use also the notation $L^p(D_\varepsilon;\delta)$ where δ is an appropriate weight function. The following statement provides some basic relations between the weighted L^p-spaces.

Lemma 3.2. *Let $1 \le p \le \infty$ and let $\varepsilon > 0$. Then $L^p(D_\varepsilon) \subset L^p(D_\varepsilon;\omega^\varepsilon)$, and for every $v \in L^p(D_\varepsilon;\omega^\varepsilon)$ there holds $v_{|D} \in L^p(D)$. Moreover,*
(i) for $\varepsilon > \tilde{\varepsilon} \ge 0$ we have

$$\|v\|_{L^p(D_{\tilde{\varepsilon}};\omega^{\tilde{\varepsilon}})} \le 2^{1/p}\|v\|_{L^p(D_\varepsilon;\omega^\varepsilon)} \quad \textit{for all } v \in L^p(D_\varepsilon;\omega^\varepsilon).$$

(ii) for $0 < \gamma < \varepsilon/2$ we have

$$\|v\|_{L^p(D_{\varepsilon-2\gamma};\omega^{\varepsilon+\gamma})} \le \left(\frac{3^\alpha\zeta_2}{\zeta_1}\right)^{1/p} \|v\|_{L^p(D_{\varepsilon-2\gamma};\omega^{\varepsilon-\gamma})} \quad \textit{for all } v \in L^p(D_\varepsilon;\omega^\varepsilon).$$

Proof. (i) For $v \in L^p(D_\varepsilon;\omega^\varepsilon)$ we obtain

$$\int_{D_{\tilde{\varepsilon}}} |v|^p\,\mathrm{d}\omega^{\tilde{\varepsilon}} = \int_{D_{\tilde{\varepsilon}}\setminus D} |v|^p\omega^{\tilde{\varepsilon}}\,\mathrm{d}x + \int_D |v|^p\omega^{\tilde{\varepsilon}}\,\mathrm{d}x.$$

For $x \in D_{\tilde{\varepsilon}} \setminus D$ we have $\omega^{\tilde{\varepsilon}}(x) \le \omega^\varepsilon(x)$, and on D there holds $1/2 \le \omega^{\tilde{\varepsilon}} \le 1 \le 2\omega^\varepsilon$. The fact that $D_{\tilde{\varepsilon}} \subset D_\varepsilon$ yields the assertion.
(ii) Similar to (i) we have $\omega^{\varepsilon+\gamma} \le \omega^{\varepsilon-\gamma}$ on D, whence

$$\|v\|^p_{L^p(D_{\varepsilon-2\gamma};\omega^{\varepsilon+\gamma})} \le \int_D |v|^p\,\mathrm{d}\omega^{\varepsilon-\gamma} + \int_{D_{\varepsilon-2\gamma}\setminus D} |v|^p\,\mathrm{d}\omega^{\varepsilon+\gamma}.$$

For $z \in D_{\varepsilon-2\gamma} \setminus D$, we can write $z = x - tn(x)$ with $x \in \partial D_{\varepsilon-2\gamma}$ and $0 \le t \le 2\varepsilon$, see Section 2.2. Then, using (12), we have

$$\omega^{\varepsilon+\gamma}(z) \le \zeta_2\left(\frac{\operatorname{dist}(z,\partial D_{\varepsilon+\gamma})}{\varepsilon+\gamma}\right)^\alpha = \zeta_2\left(\frac{t+3\gamma}{\varepsilon+\gamma}\right)^\alpha, \quad \text{and}$$
$$\omega^{\varepsilon-\gamma}(z) \ge \zeta_1\left(\frac{\operatorname{dist}(z,\partial D_{\varepsilon-\gamma})}{\varepsilon-\gamma}\right)^\alpha = \zeta_1\left(\frac{t+\gamma}{\varepsilon-\gamma}\right)^\alpha.$$

This implies

$$\frac{\omega^{\varepsilon+\gamma}(z)}{\omega^{\varepsilon-\gamma}(z)} \le \frac{\zeta_2}{\zeta_1}\left(\frac{t+3\gamma}{t+\gamma}\,\frac{\varepsilon-\gamma}{\varepsilon+\gamma}\right)^{\alpha} = \frac{\zeta_2}{\zeta_1}\left(\left(1+\frac{2\gamma}{t+\gamma}\right)\frac{\varepsilon-\gamma}{\varepsilon+\gamma}\right)^{\alpha} \le \frac{\zeta_2 3^{\alpha}}{\zeta_1},$$

from which we easily obtain the assertion. □

Associated to the weighted spaces $L^p(D_\varepsilon;\omega^\varepsilon)$, let us define the weighted Sobolev spaces

$$W^{1,p}(D_\varepsilon;\omega^\varepsilon) = \{v \in L^p(D_\varepsilon;\omega^\varepsilon) :\ \partial_{x_i} v \in L^p(D_\varepsilon;\omega^\varepsilon),\ 1 \le i \le n\}$$

with norm

$$\|v\|^p_{W^{1,p}(D_\varepsilon;\omega^\varepsilon)} = \int_{D_\varepsilon} |v|^p + |\nabla v|^p \,\mathrm{d}\omega^\varepsilon.$$

In view of (12), several results of Kufner [21] concerning power-type weights can be applied. This is due to the fact that the proofs of [21] need the power-type property of the weight only in a vicinity of the boundary; in the remaining subset of D_ε, say D, uniform boundedness away from zero and boundedness of the weight is used, which in turn allows to use results for unweighted spaces. We will employ similar techniques in the next section. We have the following. For $1 \le p < \infty$, the spaces $L^p(D_\varepsilon;\omega^\varepsilon)$ and $W^{1,p}(D_\varepsilon;\omega_\varepsilon)$ are separable Banach spaces [21, Theorem 3.6], and $C^\infty(\overline{D_\varepsilon})$ is dense in $W^{1,p}(D_\varepsilon;\omega^\varepsilon)$ [21, Theorem 7.2]. In case $p = 2$, the spaces $L^2(D_\varepsilon;\omega^\varepsilon)$ and $W^{1,2}(D_\varepsilon;\omega^\varepsilon)$ are Hilbert spaces with obvious definition of the inner product.

4 Properties of Sobolev spaces for diffuse interfaces

In this section we establish basic results for weighted Sobolev spaces for diffuse interfaces which are crucial for the analysis of variational boundary value problems. We particularly investigate the dependence on the parameter ε. As shown below, our results yield constants independent of ε, for instance the trace constant or the Poincaré constant in weighted Sobolev spaces, see Theorem 4.2 or Theorem 4.9. These results may be of interest in their own.

4.1 Trace lemma

The following auxiliary lemma is a slight adaptation of [3, Lemma 2.1]. It states that the trace operator for unweighted Sobolev spaces is uniformly bounded for certain perturbations of the domain, and it is the key observation in proving a similar statement also for weighted Sobolev spaces. For convenience of the reader, we sketch a proof.

Lemma 4.1. *Let ε_0 be sufficiently small. Then there is a constant $C > 0$ such that for each $t \in (-\varepsilon_0, \varepsilon_0)$ and $v \in W^{1,p}(D_t)$, $1 \le p < \infty$,*

$$\int_{\partial D_t} |v|^p \,\mathrm{d}\sigma \le C \int_{D_t} |\nabla v|^p + |v|^p \,\mathrm{d}x. \tag{13}$$

Proof. Let $\varepsilon_0 > 0$ be sufficiently small and let $t \in (-\varepsilon_0, \varepsilon_0)$. Let us first consider the case $p = 1$. Denote by $\Psi_t : D \to D_t$ and $\Phi_t : \partial D \to \partial D_t$ the transformations defined in Section 2.3. By a change of variables $u(x) = v(\Psi_t(x))$ for $x \in D$ and $u(x) = v(\Phi_t(x))$ for $x \in \partial D$, it follows that

$$\begin{aligned}
\inf_{x\in D} \det(D\Psi_t(x)) \int_D |\nabla u| + |u| \,\mathrm{d}x &\le \int_{D_t} |\nabla v| + |v| \,\mathrm{d}x \\
&\le \sup_{x\in D} \det(D\Psi_t(x)) \int_D |\nabla u| + |u| \,\mathrm{d}x, \\
\inf_{x\in \partial D} \det(D\Phi_t(x)) \int_{\partial D} |u| \,\mathrm{d}\sigma &\le \int_{\partial D_t} |v| \,\mathrm{d}\sigma \\
&\le \sup_{x\in \partial D} \det(D\Phi_t(x)) \int_{\partial D} |u| \,\mathrm{d}\sigma.
\end{aligned}$$

In view of Section 2.3, as $t \to 0$

$$\begin{aligned}
C_1(t) &= \min\{\inf_{x\in D} \det(D\Psi_t(x)), \inf_{x\in \partial D} \det(D\Phi_t(x))\} \to 1, \quad \text{and} \\
C_2(t) &= \max\{\sup_{x\in D} \det(D\Psi_t(x)), \sup_{x\in \partial D} \det(D\Phi_t(x))\} \to 1.
\end{aligned}$$

Denote by C a bound for the norm of the trace operator $W^{1,1}(D) \to L^1(\partial D)$. We then obtain

$$\int_{\partial D_t} |v| \,\mathrm{d}\sigma \le C \frac{C_2(t)}{C_1(t)} \int_{\partial D_t} |\nabla v| + |v| \,\mathrm{d}x.$$

For the general case $p > 1$, we apply the latter estimate to $\tilde{v} = |v|^p$. We observe that $|\nabla \tilde{v}| = p|v|^{p-1}|\nabla v|$, and, using Young's inequality, $p|v|^{p-1}|\nabla v| + |v|^p \le |\nabla v|^p + |v|^p$, which concludes the proof. □

Theorem 4.2 (Trace lemma)**.** *Let $\varepsilon_0 > 0$ be sufficiently small. Then there exists a constant $C > 0$ such that for $\varepsilon \in (0, \varepsilon_0)$ and for $v \in W^{1,p}(D_\varepsilon; \omega^\varepsilon)$, $1 \le p < \infty$,*

$$\int_{D_\varepsilon} |v|^p |\nabla \omega^\varepsilon| \,\mathrm{d}x \le C \|v\|^p_{W^{1,p}(D_\varepsilon;\omega^\varepsilon)}.$$

Proof. According to the coarea-formula there holds

$$\int_{D_\varepsilon} |v|^p |\nabla \varphi^\varepsilon| \,\mathrm{d}x = \int_{-\infty}^{\infty} \int_{(\varphi^\varepsilon)^{-1}(s)} |v|^p \,\mathrm{d}\sigma \,\mathrm{d}s = \int_{-1}^{1} \int_{\{\varphi^\varepsilon = s\}} |v|^p \,\mathrm{d}\sigma \,\mathrm{d}s.$$

Since $\{\varphi^\varepsilon = s\} = \partial\{\varphi^\varepsilon > s\} = \partial D_t$ for $t = -\varepsilon S^{-1}(s)$, we can use (13) to obtain

$$\begin{aligned}\int_{-1}^{1}\int_{\{\varphi^\varepsilon=s\}} |v|^p \,\mathrm{d}\sigma\,\mathrm{d}s &\le C\int_{-1}^{1}\int_{\{\varphi^\varepsilon>s\}} |\nabla v|^p + |v|^p \,\mathrm{d}x\,\mathrm{d}s \\ &= C\int_{D_\varepsilon}\int_{-1}^{\varphi^\varepsilon(x)} \mathrm{d}s(|\nabla v|^p + |v|^p)\,\mathrm{d}x,\end{aligned}$$

where we used Fubini's theorem in the last step. The assertion follows from $|\nabla\omega^\varepsilon| = \frac{1}{2}|\nabla\varphi^\varepsilon|$. □

The trace theorem 4.2 shows that for $g \in W^{1,p}(D_\varepsilon;\omega^\varepsilon)$ the diffuse boundary integral introduced in (5) actually exists.

4.2 Embedding theorem

The following embedding theorem uses the representation (12) of the weight near the boundary as a power of the distance function $\delta(x) = \operatorname{dist}(x, \partial D_\varepsilon)$. Let us define the Sobolev conjugate p^*_α for weighted spaces

$$p^*_\alpha = \frac{p(n+\alpha)}{n+\alpha-p} \quad \text{for } p < n+\alpha, \qquad \text{and} \qquad p^*_\alpha = \infty \quad \text{for } p \ge n+\alpha. \tag{14}$$

We observe that p^*_0 is the "usual" Sobolev conjugate for unweighted Sobolev spaces, see [1], and p^*_α is strictly decreasing with respect to α on $(0,\infty)$.

Theorem 4.3 (Embedding)**.** *Let $0 < \varepsilon < \varepsilon_0$, and let $\alpha > 0$ be the constant from (S2). Then the following embeddings are continuous*

$$W^{1,p}(D_\varepsilon,\omega^\varepsilon) \hookrightarrow L^q(D_\varepsilon,\omega^\varepsilon), \qquad 1 \le q \le p^*_\alpha \text{ and } q < \infty.$$

Moreover, there exists a constant C independent of ε such that for $u \in W^{1,p}(D_\varepsilon;\omega^\varepsilon)$

$$\|u\|_{L^q(D_\varepsilon;\omega^\varepsilon)} \le C\|u\|_{W^{1,p}(D_\varepsilon;\omega^\varepsilon)}. \tag{15}$$

The first part of the theorem can be found in [20, Theorem 3], see also [28, Theorem 19.9] for the case $q < p^*_\alpha$. To show that the embedding is independent of ε, we will give a proof in the spirit of [28]. To do so, we employ the following two lemmata. The first of which uses Sobolev's embedding theorem on balls and a covering argument, and is similar to the arguments of [28]. The second is a Hardy-inequality-type argument for diffuse interfaces which seems to be new. We let $\delta(x) = \operatorname{dist}(x, \partial D_\varepsilon)$ in the following.

Lemma 4.4. *Let $\varepsilon > 0$ and $\alpha > 0$. Furthermore, let $1 \le q < \infty$ such that $n + \alpha \ge (n+\alpha-1)q$. Then there exists a constant $C > 0$ independent of ε such that for every $u \in W^{1,1}(D_\varepsilon;\delta^\alpha)$*

$$\|u\|_{L^q(D_\varepsilon\setminus D;\delta^\alpha)} \le C\left(\|u\|_{L^1(\Gamma_\varepsilon;\delta^{\alpha-1})} + \|\nabla u\|_{L^1(\Gamma_\varepsilon;\delta^\alpha)}\right).$$

Proof. We proceed as in [28]. Let $r(x) = \delta(x)/3$. According to the Besicovitch covering theorem, cf. [28, Lemma 18.3], there exists a sequence $\{x_k\} \subset D_\varepsilon \setminus D$ and an integer θ depending only on n such that

$$D_\varepsilon \setminus D \subset \bigcup_{k=1}^{\infty} B_k \subset \Gamma_\varepsilon, \quad B_k = B_{r(x_k)}(x_k) \text{ and } \sum_{k=1}^{\infty} \chi_{B_k}(x) \leq \theta \text{ for all } x \in \mathbb{R}^n.$$

Employing the Sobolev embedding theorem for balls [1], we obtain as in [28, Theorem 18.6]

$$\|u\|_{L^q(B_k;\delta^\alpha)} \leq C\delta(x_k)^{\frac{n+\alpha}{q}-n+1-\alpha} \left(\|u\|_{L^1(B_k;\delta^{\alpha-1})} + \|\nabla u\|_{L^1(B_k;\delta^\alpha)}\right) \quad (16)$$

for all $q \leq n/(n-1)$. Note the different powers of the weight on the right-hand side of (16). Assuming, without loss of generality, that $\varepsilon_0 \leq 1$ and thus $\delta(x_k) \leq 1$, we can bound the right-hand side of (16) as long as α and q are such that $n + \alpha \geq (n + \alpha - 1)q$. Hence, by summation over k, we obtain the assertion with a constant C depending on n and the Sobolev embedding constant for the unit ball, but not on ε. □

Lemma 4.5 (Hardy-type inequality)**.** *Let $0 < \varepsilon < \varepsilon_0$ and let $\alpha > 0$. Then there exists a constant $C > 0$ independent of ε such that for every $u \in W^{1,1}(D_\varepsilon; \delta^\alpha)$*

$$\|u\|_{L^1(\Gamma_\varepsilon;\delta^{\alpha-1})} \leq \frac{C}{\alpha} \left(\varepsilon^\alpha \|u\|_{W^{1,1}(D)} + \|\nabla u\|_{L^1(\Gamma_\varepsilon;\delta^\alpha)}\right).$$

Proof. Let $u \in C^\infty(\overline{D_\varepsilon}) \cap W^{1,1}(D_\varepsilon; \delta^\alpha)$. We obtain by using (11), $1/2 \leq |\det D\Phi_{-t}| \leq 2$, and one-dimensional integration-by-parts

$$\begin{aligned}
\int_{\Gamma_\varepsilon} |u|\delta^{\alpha-1} \,\mathrm{d}x &= \int_{\partial D_\varepsilon} \int_0^{2\varepsilon} |u(x - tn(x))| t^{\alpha-1} |\det D\Phi_{-t}| \,\mathrm{d}t \,\mathrm{d}\sigma_\varepsilon(x) \\
&\leq \frac{2}{\alpha} \int_{\partial D_\varepsilon} |u(x - 2\varepsilon n(x))| (2\varepsilon)^\alpha \,\mathrm{d}\sigma_\varepsilon(x) \\
&\quad + \frac{2}{\alpha} \int_{\partial D_\varepsilon} \int_0^{2\varepsilon} |\nabla u(x - tn(x))| t^\alpha dt \,\mathrm{d}\sigma_\varepsilon(x). \qquad (17)
\end{aligned}$$

To treat the first integral in (17), we employ the transformation $\Phi_{2\varepsilon} : \partial D_{-\varepsilon} \to \partial D_\varepsilon$ defined in Section 2.3, i.e.

$$\int_{\partial D_\varepsilon} |u(x - 2\varepsilon n(x))| \,\mathrm{d}\sigma_\varepsilon(x) \leq 4 \int_{\partial D_{-\varepsilon}} |u(x)| \,\mathrm{d}\sigma_{-\varepsilon}(x).$$

From $u_{|D} \in W^{1,1}(D)$, and $D_{-\varepsilon} \subset D$, and the trace lemma 4.1 we deduce that there exists a constant $C > 0$ independent of ε such that

$$\int_{\partial D_{-\varepsilon}} |u(x)| \,\mathrm{d}\sigma_{-\varepsilon}(x) \leq C\|u\|_{W^{1,1}(D)},$$

i.e. $\int_{\partial D_\varepsilon} |u(x-2\varepsilon n(x))|\,\mathrm{d}\sigma_\varepsilon(x) \le C\|u\|_{W^{1,1}(D)}$. The assertion follows from

$$\int_{\partial D_\varepsilon}\int_0^{2\varepsilon} |\nabla u(x-tn(x))| t^\alpha \,\mathrm{d}t\,\mathrm{d}\sigma_\varepsilon(x) \le 2\int_{\Gamma_\varepsilon} |\nabla u|\delta^\alpha \,\mathrm{d}x,$$

and a density argument. □

Remark 4.6. *Let us state that the arguments of [21] are based on partition of unity $\{\varphi_i\}$ subordinate to $\{B_j\}\cup\{D\}$ where $\{B_j\}$ is a finite cover of Γ_ε. Then the Hardy-type argument of the latter proof is applied to $v_i = u\varphi_i$ which is zero on the boundary of B_i. Hence, the first term in* (17) *vanishes. However, $\nabla v_i = \varphi_i\nabla u + u\nabla\varphi_i$, and $|\nabla\varphi_i| \sim 1/\varepsilon$. Thus, the techniques of [21] are not directly applicable as we strive for constants uniformly bounded in terms of ε.*

Proof of Theorem 4.3. We split the norm into the diffuse interface part and the interior part

$$\|u\|^q_{L^q(D_\varepsilon;\omega^\varepsilon)} = \int_{D_\varepsilon\setminus D} |u|^q\omega^\varepsilon \,\mathrm{d}x + \int_D |u|^q\omega^\varepsilon\,\mathrm{d}x.$$

From the Sobolev embedding theorem [1], we have that $W^{1,p}(D;\omega^\varepsilon) \hookrightarrow L^q(D;\omega^\varepsilon)$ is continuous for each $q \le p_0^* = np/(n-p)$ if $p < n$, and for $q < \infty$ if $p \ge n$. Since $p_\alpha^* \le p_0^*$ and D is bounded, we only have to estimate the L^q-norm of u on $D_\varepsilon \setminus D$.
First consider the case $p = 1$, $q \le p_\alpha^* = (n+\alpha)/(n+\alpha-1)$. For this choice, the condition $n+\alpha \ge (n+\alpha-1)q$ is obviously satisfied. Combining Lemma 4.5 and Lemma 4.4 yields

$$\begin{aligned}\int_{D_\varepsilon\setminus D} |u|^q\delta^\alpha\,\mathrm{d}x &\le C\left(\int_{\Gamma_\varepsilon} |u|\delta^{\alpha-1} + |\nabla u|\delta^\alpha\,\mathrm{d}x\right)^q \\ &\le C\left(\varepsilon^\alpha\|u\|_{W^{1,1}(D)} + \|\nabla u\|_{L^1(\Gamma_\varepsilon;\delta^\alpha)}\right)^q.\end{aligned}$$

Multiplication of the latter inequality with $1/\varepsilon^\alpha$, taking the qth root and using (12), i.e. $\delta^\alpha/\varepsilon^\alpha \approx \omega^\varepsilon$, further gives,

$$\|u\|_{L^q(D_\varepsilon\setminus D;\omega^\varepsilon)} \le C\varepsilon^{\alpha(1-1/q)}\left(\|u\|_{W^{1,1}(D)} + \|\nabla u\|_{L^1(\Gamma_\varepsilon;\omega^\varepsilon)}\right).$$

Summarizing, we have shown that for each $1 \le q \le (n+\alpha)/(n+\alpha-1)$ there holds

$$\|u\|_{L^q(D_\varepsilon;\omega^\varepsilon)} \le C\|u\|_{W^{1,1}(D_\varepsilon;\omega^\varepsilon)}.$$

For the general case $p > 1$, we apply the previous results to $v = |u|^{1+q(p-1)/p}$ and $\tilde q = (1 - \frac{1}{p} + \frac{1}{q})^{-1}$. One easily verifies that $q \le p(n+\alpha)/(n+\alpha-p)$

is equivalent to $\tilde{q} \le (n+\alpha)/(n+\alpha-1)$. Moreover, $|v|^{\tilde{q}} = |u|^q$ and $|\nabla v| = (1+\frac{q(p-1)}{p})|u|^{q(p-1)/p}|\nabla u|$. Whence, Hölder's inequality yields

$$\begin{aligned}\|v\|_{W^{1,1}(D_\varepsilon;\omega^\varepsilon)} &= \int_{D_\varepsilon} |u|^{1+\frac{q(p-1)}{p}} + \big(1+\frac{q(p-1)}{p}\big)|u|^{\frac{q(p-1)}{p}}|\nabla u|\,\mathrm{d}\omega^\varepsilon \\ &\le \left(\|u\|_{L^p(D_\varepsilon;\omega^\varepsilon)} + \big(1+\frac{q(p-1)}{p}\big)\|\nabla u\|_{L^p(D_\varepsilon;\omega^\varepsilon)}\right)\|u\|^{\frac{q(p-1)}{p}}_{L^q(D_\varepsilon;\omega^\varepsilon)}.\end{aligned}$$

This together with the identity

$$\|v\|_{L^{\tilde{q}}(D_\varepsilon;\omega^\varepsilon)} = \|u\|^{1+\frac{q(p-1)}{p}}_{L^q(D_\varepsilon;\omega^\varepsilon)}$$

yields the assertion. □

Remark 4.7. *As already noted, p^*_α is strictly decreasing with respect to α on $(0,\infty)$. Loosely speaking, compared to the unweighted Sobolev embedding, we loose α spatial dimensions. For instance, if $\alpha = 1$, we have that $2^*_1 = 6$ for $n = 2$, and $2^*_1 = 4$ for $n = 3$. This fact is intimately related to the Hardy inequality and isoperimetric inequalities, cf. [20] where also counterexamples are given showing that the restriction $q \le p^*_\alpha$ cannot be improved in general. However, embedding in certain Hölder spaces is possible [20, 28]. Adapting the above proofs it should be possible to show that even in this situation the embedding constants are independent of ε.*

Proposition 4.8 (Compactness)**.** *Let $0 < \varepsilon < \varepsilon_0$, α be the constant in (S2), and let $1 \le p < \infty$. Then the following embeddings are compact*

$$W^{1,p}(D_\varepsilon,\omega^\varepsilon) \hookrightarrow L^q(D_\varepsilon,\omega^\varepsilon), \qquad 1 \le q < p^*_\alpha.$$

Proof. Let $q < p^*_\alpha$ and let $\{u_k\} \subset W^{1,p}(D_\varepsilon;\omega^\varepsilon)$ be bounded; say by a constant $C_p > 0$. Furthermore denote by C_e the constant of the embedding $W^{1,p}(D_\varepsilon;\omega^\varepsilon) \to L^{p^*_\alpha}(D_\varepsilon;\omega^\varepsilon)$. Since $L^q(D_\varepsilon;\omega^\varepsilon)$ is complete, we have to show that a subsequence of $\{u_k\}$ is Cauchy in $L^q(D_\varepsilon;\omega^\varepsilon)$. Therefore, let $\iota > 0$ and choose $\gamma = \min\{\varepsilon, (C_pC_e)^{qp/(q-p^*_\alpha)}\iota/2\}$. Since the embedding $W^{1,p}(D_{\varepsilon-\gamma};\omega^\varepsilon) \hookrightarrow L^q(D_{\varepsilon-\gamma};\omega^\varepsilon)$ is compact [1], we can extract a subsequence, again denoted by $\{u_k\}$, which is Cauchy in $L^q(D_{\varepsilon-\gamma};\omega^\varepsilon)$. Hence, there exists $N = N(\iota) \in \mathbb{N}$ such that $\|u_k - u_l\|_{L^q(D_{\varepsilon-\gamma};\omega^\varepsilon)} < \iota/2$ for all $k,l \ge N$. Let $k,l \ge N$ in the following. Thus, using the triangle inequality, we have that

$$\|u_k - u_l\|_{L^q(D_\varepsilon;\omega^\varepsilon)} \le \|u_k - u_l\|_{L^q(D_\varepsilon\setminus D_{\varepsilon-\gamma};\omega^\varepsilon)} + \frac{\iota}{2}. \tag{18}$$

Since $1 \le q < p^*_\alpha$, we obtain by using Hölder's inequality and the embedding theorem

$$\begin{aligned}\|u_k - u_l\|_{L^q(D_\varepsilon\setminus D_{\varepsilon-\gamma};\omega^\varepsilon)} &\le \|u_k - u_l\|_{L^{p^*_\alpha}(D_\varepsilon\setminus D_{\varepsilon-\gamma};\omega^\varepsilon)}\gamma^{\frac{1}{q}-\frac{1}{p^*_\alpha}} \\ &\le C_e\|u_k - u_l\|_{W^{1,p}(D_\varepsilon;\omega^\varepsilon)}\gamma^{\frac{1}{q}-\frac{1}{p^*_\alpha}} \le \frac{\iota}{2}\end{aligned}$$

by choice of γ. This in combination with (18) shows that $\{u_k\}$ is Cauchy in $L^q(D_\varepsilon;\omega^\varepsilon)$. □

The idea of the proof of the previous compactness result can already be found in [28].

4.3 Diffuse Poincaré-Friedrichs inequalities

The last issue concerning basic results in weighted Sobolev spaces are Poincaré-Friedrichs inequalities, which we again want to derive with constants independent of ε. We start with a quite general result:

Theorem 4.9 (Poincaré-type inequality)**.** *Fix $1 \le p < \infty$. Assume that D_ε is connected for each $\varepsilon \in [0,\varepsilon_0]$, and let $K_\varepsilon \subset W^{1,p}(D_\varepsilon;\omega^\varepsilon)$, be a family of closed cones, i.e. for $u \in K_\varepsilon$ there holds $\lambda u \in K_\varepsilon$ for all $\lambda > 0$, such that K_ε contains only the zero function as a constant function. Then there exists a constant $C > 0$ independent of ε such that*

$$\|u\|_{L^p(D_\varepsilon;\omega^\varepsilon)} \le C\|\nabla u\|_{L^p(D_\varepsilon;\omega^\varepsilon)} \quad \text{for all } u \in K_\varepsilon. \tag{19}$$

Proof. Assume (19) is not true. Then there exist sequences $\{u_k\} \subset W^{1,p}(D_{\varepsilon_k};\omega^{\varepsilon_k}) \cap K_{\varepsilon_k}$ with $\|u_k\|_{L^p(D_{\varepsilon_k};\omega^{\varepsilon_k})} = 1$ and $\{\varepsilon_k\} \subset [0,\varepsilon_0]$ such that

$$\|\nabla u_k\|_{L^p(D_{\varepsilon_k};\omega^{\varepsilon_k})} \le \frac{1}{k}. \tag{20}$$

Since $\varepsilon_k \in [0,\varepsilon_0]$ the Bolzano-Weierstraß theorem implies the existence of a $\tilde\varepsilon \in [0,\varepsilon_0]$ such that for a subsequence, relabeled if necessary, $\varepsilon_k \to \tilde\varepsilon$ as $k \to \infty$. Hence, for all $\gamma > 0$ there exists $N(\gamma) \in \mathbb{N}$ such that $\varepsilon_k \in (\tilde\varepsilon - \gamma, \tilde\varepsilon + \gamma)$ for all $k \ge N(\gamma)$. In the following let $0 < \gamma < \tilde\varepsilon/2$ and $k \ge N(\gamma)$. By Hölder's inequality and the embedding theorem for $q = p^*_\alpha$, we have

$$\begin{aligned}\|u_k\|_{L^p(D_{\varepsilon_k}\setminus D_{\tilde\varepsilon-2\gamma};\omega^{\varepsilon_k})} &\le \|u_k\|_{L^q(D_{\varepsilon_k}\setminus D_{\tilde\varepsilon-2\gamma};\omega^{\varepsilon_k})}(2\gamma)^{\frac{1}{n+\alpha}} \\ &\le C\|u_k\|_{W^{1,p}(D_{\varepsilon_k};\omega^{\varepsilon_k})}\gamma^{\frac{1}{n+\alpha}}. \end{aligned}\tag{21}$$

Furthermore, since $\tilde\varepsilon - \gamma \le \varepsilon_k$, by Lemma 3.2 (i)

$$\|u_k\|_{W^{1,p}(D_{\tilde\varepsilon-\gamma};\omega^{\tilde\varepsilon-\gamma})} \le \|u_k\|_{W^{1,p}(D_{\varepsilon_k};\omega^{\varepsilon_k})} \le C.$$

In view of Proposition 4.8, we can therefore extract a subsequence, relabeled if necessary, such that $u_k \to u$ in $L^p(D_{\tilde\varepsilon-\gamma};\omega^{\tilde\varepsilon-\gamma})$. Moreover, we deduce from (20) and Lemma 3.2 (i) that

$$\|\nabla u_k\|_{L^p(D_{\tilde\varepsilon-\gamma};\omega^{\tilde\varepsilon-\gamma})} \le \|\nabla u_k\|_{L^p(D_{\varepsilon_k};\omega^{\varepsilon_k})} \le \frac{1}{k},$$

and hence $u_k \to u$ in $W^{1,p}(D_{\tilde{\varepsilon}-\gamma};\omega^{\tilde{\varepsilon}-\gamma})$ and $\nabla u = 0$. Since $K_{\tilde{\varepsilon}-\gamma}$ is closed and $D_{\tilde{\varepsilon}-\gamma}$ is connected, $u \in K_{\tilde{\varepsilon}-\gamma}$ and $u = 0$ on $D_{\tilde{\varepsilon}-\gamma}$. In view of Lemma 3.2 (ii), we further obtain

$$\begin{aligned}\|u_k\|_{L^p(D_{\tilde{\varepsilon}-2\gamma};\omega^{\varepsilon_k})} \le 2\|u_k\|_{L^p(D_{\tilde{\varepsilon}-2\gamma};\omega^{\tilde{\varepsilon}+\gamma})} &\le C\|u_k\|_{L^p(D_{\tilde{\varepsilon}-2\gamma};\omega^{\tilde{\varepsilon}-\gamma})} \\ &\le C\|u_k\|_{L^p(D_{\tilde{\varepsilon}-\gamma};\omega^{\tilde{\varepsilon}-\gamma})}.\end{aligned}$$

This in combination with (21) implies

$$1 = \lim_{k\to\infty} \|u_k\|_{L^p(D_{\varepsilon_k};\omega^{\varepsilon_k})} \le C\gamma^{\frac{1}{n+\alpha}} + \lim_{k\to\infty} \|u_k\|_{L^p(D_{\tilde{\varepsilon}-2\gamma};\omega^{\varepsilon_k})} = C\gamma^{\frac{1}{n+\alpha}}.$$

Since $\gamma > 0$ was arbitrary, this is the desired contradiction. □

Let us remark that in [8] a similar result has been obtained for unweighted spaces, i.e. a Poincaré inequality with a constant which is independent of certain perturbations of ∂D. For illustration of the previous result let us state the "usual" Poincaré and Friedrichs inequality in their weighted form.

Corollary 4.10. *Let $\varepsilon \in [0, \varepsilon_0]$, $1 \le p < \infty$, and let D_ε be connected. Then there exists a constant C independent of ε such that*

$$\|u - \bar{u}_{D_\varepsilon}\|_{L^p(D_\varepsilon;\omega^\varepsilon)} \le C\|\nabla u\|_{L^p(D_\varepsilon;\omega^\varepsilon)} \quad \text{for all } u \in W^{1,p}(D_\varepsilon;\omega^\varepsilon).$$

Here $\bar{u}_{D_\varepsilon} = \int_{D_\varepsilon} u \,\mathrm{d}\omega^\varepsilon / \|1\|_{L^1(D_\varepsilon;\omega^\varepsilon)}$ is the weighted mean value.

Proof. Define $K_\varepsilon = \{u \in W^{1,p}(D_\varepsilon;\omega^\varepsilon) : \bar{u}_{D_\varepsilon} = 0\}$ and use Theorem 4.9. □

Corollary 4.11 (Poincaré-Friedrichs-type inequality)**.** *Let $\varepsilon \in [0, \varepsilon_0]$, $1 \le p < \infty$, and let D_ε be connected. Then there exists a constant C independent of ε such that for every $\varepsilon \in (0, \varepsilon_0)$ and $v \in W^{1,p}(D_\varepsilon;\omega^\varepsilon)$ there holds*

$$\|v\|^p_{L^p(D_\varepsilon;\omega^\varepsilon)} \le C_P \left(\|\nabla v\|^p_{L^p(D_\varepsilon;\omega^\varepsilon)} + \int_{D_\varepsilon} |v|^p |\nabla \omega^\varepsilon| \,\mathrm{d}x \right).$$

Proof. Define $K_\varepsilon = \{v \in W^{1,p}(D_\varepsilon;\omega^\varepsilon) : \int_{D_\varepsilon} |v|^p |\nabla\omega^\varepsilon| \,\mathrm{d}x = 0\}$ and use Theorem 4.9. □

Remark 4.12. *In Corollary 4.11 one can make the constant explicit if one applies the "classical" Poincaré-Friedrichs inequality [1, 6.26] to $|v|^p\omega^\varepsilon \in W^{1,1}_0(\Omega)$.*

Remark 4.13. *Theorem 4.9 also holds for the case $p = \infty$: By Rellich's theorem [1] the embedding $W^{1,\infty}(D_{\tilde{\varepsilon}-\gamma}) \hookrightarrow C^{0,1}(\overline{D_{\tilde{\varepsilon}-\gamma}}) \hookrightarrow L^\infty(D_{\tilde{\varepsilon}-\gamma})$ is compact. Then, with similar arguments as above, the assumption (20) with $p = \infty$ leads to $\|u_k\|_{L^\infty(D_{\tilde{\varepsilon}-\gamma})} \to 0$. Then, for $\tilde{x} = x + tn(x) \in D_{\varepsilon_k} \setminus D_{\tilde{\varepsilon}-\gamma}$ with $x \in \partial D_{\tilde{\varepsilon}-\gamma}$ and $t \le \gamma$, we obtain as $k \to \infty$*

$$|u_k(\tilde{x})| \le |u_k(x)| + \gamma\|\nabla u_k\|_{L^p(D_{\varepsilon_k})} \le |u_k(x)| + \gamma/k \to 0,$$

where we have chosen a Lipschitz continuous representative of u_k. Hence, $\|u_k\|_{L^\infty(D_{\varepsilon_k})} \to 0$ which contradicts $\|u_k\|_{L^\infty(D_{\varepsilon_k})} = 1$.

5 Convergence of diffuse integrals

In the following two subsections we investigate the approximation properties of the diffuse integrals introduced in (5).

5.1 Convergence of diffuse volume integrals

We start with the case of volume integrals, for which we want to estimate the error

$$E_V = \int_\Omega h(x)\,\mathrm{d}\omega^\varepsilon(x) - \int_D h(x)\,\mathrm{d}x$$

between the volume integral and the diffuse volume integral in terms of ε and h. We will provide estimates for the cases $h \in L^p(D_\varepsilon;\omega^\varepsilon)$ and $h \in W^{1,p}(D_\varepsilon;\omega^\varepsilon)$ which gives stronger results improved by one order of ε.

Theorem 5.1. *Let $1 < p \leq \infty$ and $h \in L^p(D_\varepsilon;\omega^\varepsilon)$. Then there exists a constant $C > 0$ independent of ε such that*

$$|E_V| \leq C\varepsilon^{1-\frac{1}{p}}\|h\|_{L^p(\Gamma_\varepsilon;\omega^\varepsilon)}.$$

Moreover, if $p = 1$ and $h \in L^1(\Omega)$, then $E_V \to 0$ as $\varepsilon \to 0^+$.

Proof. L^1-regularity: Let $h \in L^1(\Omega)$. Using dominated convergence, we infer from $\omega^\varepsilon(x) \to \chi_D(x)$ as $\varepsilon \to 0^+$ a.e. $x \in \Omega$ and $h\omega^\varepsilon \leq h$ that

$$\lim_{\varepsilon\to 0^+} E_V = 0.$$

Let $h \in L^p(D_\varepsilon;\omega^\varepsilon)$ for fixed but arbitrary $1 < p \leq \infty$. Using (7) and $\omega^\varepsilon = 1$ on $D \setminus \Gamma_\varepsilon$, we obtain the representation

$$E_V = \int_{\Gamma_\varepsilon} h(x)\,\mathrm{d}\omega^\varepsilon(x) - \int_{D\cap\Gamma_\varepsilon} h(x)\,\mathrm{d}x.$$

Using Hölders inequality and $1 \leq 2\omega^\varepsilon$ on $D \cap \Gamma_\varepsilon$ we can estimate the two terms as follows

$$\int_{\Gamma_\varepsilon} h(x)\,\mathrm{d}\omega^\varepsilon(x) \leq \|h\|_{L^p(\Gamma_\varepsilon;\omega^\varepsilon)}\|\omega^\varepsilon\|^{1-\frac{1}{p}}_{L^1(\Gamma_\varepsilon)},$$
$$\int_{D\cap\Gamma_\varepsilon} |h(x)|\,\mathrm{d}x \leq 2\int_{D\cap\Gamma_\varepsilon} |h(x)|\omega^\varepsilon(x)\,\mathrm{d}x \leq 2\|h\|_{L^p(\Gamma_\varepsilon;\omega^\varepsilon)}\|\omega^\varepsilon\|^{1-\frac{1}{p}}_{L^1(D\cap\Gamma_\varepsilon)}.$$

Since $|\omega^\varepsilon| \leq 1$, we deduce from (8) that

$$\|\omega^\varepsilon\|^{1-\frac{1}{p}}_{L^1(\Gamma_\varepsilon)} \leq C\varepsilon^{1-\frac{1}{p}},$$

which concludes the proof. □

Theorem 5.1 for L^p-functions relies basically on the fact that $|\Gamma_\varepsilon| \leq C\varepsilon$. This is due to the fact that $L^p(D_\varepsilon;\omega^\varepsilon)$-functions can have singularities in Γ_ε. Note that in the case $p = 1$ we expect no rate of convergence in terms of ε, and the assumption $h \in L^1(\Omega)$ is stronger than those for $p > 1$. Resorting to $W^{1,p}$-functions we can exploit extensively symmetry of the phase-field function ω^ε leading to a much stronger result.

Theorem 5.2. *Let $0 < \varepsilon \leq \varepsilon_0$, and let $h \in W^{1,p}(D_\varepsilon;\omega^\varepsilon)$ for some $1 \leq p \leq \infty$. Then there exists $C > 0$ independent of ε such that*

$$|E_V| \leq C\varepsilon^{2-\frac{1}{p}}\|h\|_{W^{1,p}(\Gamma_\varepsilon;\omega^\varepsilon)}.$$

Proof. Using a change of variables $s = S(-t/\varepsilon)$ and Fubini's theorem, we observe that

$$\int_{-\varepsilon}^{\varepsilon} \frac{1}{2\varepsilon} S'(-\frac{t}{\varepsilon}) \int_{\{d_D(x)<t\}} h(x)\,\mathrm{d}x\,\mathrm{d}t = \int_\Omega h(x)\frac{1+\varphi^\varepsilon}{2}\,\mathrm{d}x.$$

Since $\int_{-\varepsilon}^{\varepsilon} \frac{1}{2\varepsilon} S'(-\frac{t}{\varepsilon})\,\mathrm{d}t = 1$, we further obtain

$$E_V = \int_{-\varepsilon}^{\varepsilon} \frac{1}{2\varepsilon} S'(-\frac{t}{\varepsilon}) \left(\int_{\{d_D(x)<t\}} h(x)\,\mathrm{d}x - \int_{\{d_D(x)<0\}} h(x)\,\mathrm{d}x \right) \mathrm{d}t.$$

Observing that

$$\int_{\{d_D(x)<t\}} h(x)\,\mathrm{d}x - \int_{\{d_D(x)<0\}} h(x)\,\mathrm{d}x = -\int_{\{t<d_D(x)<0\}} h(x)\,\mathrm{d}x \quad \text{for } t < 0$$

and

$$\int_{\{d_D(x)<t\}} h(x)\,\mathrm{d}x - \int_{\{d_D(x)<0\}} h(x)\,\mathrm{d}x = \int_{\{0<d_D(x)<t\}} h(x)\,\mathrm{d}x \quad \text{for } t > 0,$$

and splitting the integration over $(-\varepsilon,\varepsilon)$ to $(-\varepsilon,0)$ and $(0,\varepsilon)$ and employing a change of variables $t \mapsto -t$ for the integral over $(-\varepsilon,0)$, we further obtain

$$E_V = \int_0^{\varepsilon} \frac{1}{2\varepsilon} S'(-\frac{t}{\varepsilon}) \left(\int_{\{0<d_D(x)<t\}} h(x)\,\mathrm{d}x - \int_{\{-t<d_D(x)<0\}} h(x)\,\mathrm{d}x \right) \mathrm{d}t. \tag{22}$$

For the last computation, we used $S(-t) = -S(t)$, i.e. $S'(-t) = S'(t)$. To compare the difference on the right-hand side of the latter equation we use the transformations Φ_s introduced in Section 2.3, and the transformation formula, namely

$$\begin{aligned}
&\int_{\{0<d_D(x)<t\}} h(x)\,\mathrm{d}x - \int_{\{-t<d_D(x)<0\}} h(x)\,\mathrm{d}x \\
&= \int_0^t \int_{\partial D} \big(h(x+sn(x)) - h(x-sn(x))\big)|\det D\Phi_s(x)|\,\mathrm{d}\sigma(x)\,\mathrm{d}s \\
&\quad + \int_0^t \int_{\partial D} h(x-sn(x))(|\det D\Phi_s(x)| - |\det D\Phi_{-s}(x)|)\,\mathrm{d}\sigma(x)\,\mathrm{d}s.
\end{aligned}$$

The two integrals can be treated separately. Using $h(x+sn(x))-h(x-sn(x))=\int_{-s}^{s}\nabla h(x+\tau n(x))\cdot n(x)d\tau$, and $\frac{1}{2}\le|\det D\Phi_s(x)|\le 2$ we obtain using Fubini's theorem and the transformation formula

$$\begin{aligned}
&\int_0^t\int_{\partial D}\big(h(x+sn(x))-h(x-sn(x))\big)|\det D\Phi_s(x)|\,\mathrm{d}\sigma(x)\,\mathrm{d}s\\
&\le 2\int_0^t\int_{-s}^{s}\int_{\partial D}|\nabla h(x+\tau n(x))\cdot n(x)|\,\mathrm{d}\sigma(x)d\tau\,\mathrm{d}s\\
&\le 4t\int_{-t}^{t}\int_{\partial D}|\nabla h(x+\tau n(x))||\det D\Phi_\tau(x)|\,\mathrm{d}\sigma(x)d\tau\\
&= 4t\int_{\Gamma_t}|\nabla h(x)|\,\mathrm{d}x.
\end{aligned}$$

For the second integral we obtain

$$\begin{aligned}
&\int_0^t\int_{\partial D}h(x-sn(x))(|\det D\Phi_s(x)|-|\det D\Phi_{-s}(x)|)\,\mathrm{d}\sigma(x)\,\mathrm{d}s\\
&\qquad\le Ct\int_{\Gamma_t}|h(x)|\,\mathrm{d}x \qquad (23)
\end{aligned}$$

which can be seeen with (10) as

$$\begin{aligned}
&\int_0^t\int_{\partial D}h(x-sn(x))(|\det D\Phi_s(x)|-|\det D\Phi_{-s}(x)|)\,\mathrm{d}\sigma(x)\,\mathrm{d}s\\
&\le C\|D^2d_D\|_{L^\infty(\partial D)}\int_0^t s\int_{\partial D}|h(x-sn(x))|\,\mathrm{d}\sigma(x)\,\mathrm{d}s\\
&\le 2C\|D^2d_D\|_{L^\infty(\partial D)}t\int_0^t\int_{\partial D}|h(x-sn(x))||\det D\Phi_{-s}(x)|\,\mathrm{d}\sigma(x)\,\mathrm{d}s\\
&\le 2C\|D^2d_D\|_{L^\infty(\partial D)}t\int_{\Gamma_t}|h(x)|\,\mathrm{d}x.
\end{aligned}$$

Using these estimates we obtain from (22)

$$|E_V|\le C\int_0^\varepsilon\frac{1}{2\varepsilon}S'(-\frac{t}{\varepsilon})t\int_{\Gamma_t}|h(x)|+|\nabla h(x)|\,\mathrm{d}x\,\mathrm{d}t.$$

Setting $p'=p/(p-1)$, an application of Hölder's inequality thus yields

$$\begin{aligned}
|E_V|\le\frac{C}{2\varepsilon}&\left(\int_0^\varepsilon S'(-\frac{t}{\varepsilon})t^{p'}\,\mathrm{d}t\right)^{\frac{1}{p'}}\\
&\left(\int_0^\varepsilon S'(-\frac{t}{\varepsilon})\left(\int_{\Gamma_t}|h(x)|+|\nabla h(x)|\,\mathrm{d}x\right)^p\,\mathrm{d}t\right)^{\frac{1}{p}}.
\end{aligned}$$

Using boundedness of S' the first integral can be computed explicitly. To treat the second integral, we use Hölder's inequality for the inner integral which gives

$$|E_V| \leq \frac{C}{\varepsilon}\varepsilon^{2-\frac{1}{p}} \left(\int_0^\varepsilon S'(-\frac{t}{\varepsilon})|\Gamma_t|^{\frac{p}{p'}} \int_{\Gamma_t} |h(x)|^p + |\nabla h(x)|^p \,\mathrm{d}x \,\mathrm{d}t \right)^{\frac{1}{p}}.$$

Note, that C is a universal constant depending only on S, D and p but not on ε or h which may change from to line. Using $|\Gamma_t| \leq Ct|\partial D|$, $t \leq \varepsilon$ and $\frac{1}{p} + \frac{1}{p'} = 1$, we therefore have

$$|E_V| \leq C\varepsilon^{2-\frac{1}{p}} \left(\frac{1}{\varepsilon} \int_0^\varepsilon S'(-\frac{t}{\varepsilon}) \int_{\Gamma_t} |h(x)|^p + |\nabla h(x)|^p \,\mathrm{d}x \,\mathrm{d}t \right)^{\frac{1}{p}}.$$

Since $\Gamma_t = \{x \in D_\varepsilon : \ -t < d_D(x) < t\} = \{x \in D_\varepsilon : -s < \varphi^\varepsilon(x) < s\}$ for $s = -S(-\frac{t}{\varepsilon})$, a corresponding transformation yields

$$\begin{aligned}
&\frac{1}{\varepsilon} \int_0^\varepsilon S'(-\frac{t}{\varepsilon}) \int_{\Gamma_t} |h(x)|^p + |\nabla h(x)|^p \,\mathrm{d}x \,\mathrm{d}t \\
&\qquad = \int_0^1 \int_{\{-s<\varphi^\varepsilon<s\}} |h(x)|^p + |\nabla h(x)|^p \,\mathrm{d}x \,\mathrm{d}s \\
&\qquad = \int_{\Gamma_\varepsilon} \int_{|\varphi^\varepsilon(x)|}^1 \mathrm{d}s \big(|h(x)|^p + |\nabla h(x)|^p \big) \,\mathrm{d}x \leq 2\|h\|^p_{W^{1,p}(\Gamma_\varepsilon;\omega^\varepsilon)}
\end{aligned}$$

where we used that $1 - |\varphi^\varepsilon| \leq 2\omega^\varepsilon$ on Γ_ε, and

$$\begin{aligned}
&\{(x,s) \in \mathbb{R}^{n+1} : 0 < s < 1, \ -s < \varphi^\varepsilon(x) < s\} \\
&\qquad = \{(x,s) \in \mathbb{R}^{n+1} : |\varphi^\varepsilon(x)| < s < 1\}.
\end{aligned}$$

This yields the assertion. □

For sake of completeness, let us state a corresponding approximation result for Hölder continuous function, i.e. we say that $h \in C^{0,\nu}(\overline{\Omega})$, if h is continuous on $\overline{\Omega}$ and if

$$|h|_\nu = \sup_{x \neq y} \frac{h(x) - h(y)}{|x - y|^\nu} < \infty.$$

We write $\|h\|_{C^{0,\nu}(\overline{\Omega})} = \sup_{x\in\Omega} |h(x)| + |h|_\nu$.

Lemma 5.3. *Let $0 < \varepsilon \leq \varepsilon_0$, and let $h \in C^{0,\nu}(\overline{\Gamma_\varepsilon})$ for some $0 < \nu \leq 1$. Then there exists $C > 0$ independent of ε such that*

$$|E_V| \leq C\|h\|_{C^{0,\nu}(\overline{\Gamma_\varepsilon})}\varepsilon^{\nu+1}.$$

Proof. It is easy to show the estimates

$$\int_0^t \int_{\partial D} \big(h(x+sn(x)) - h(x-sn(x))\big) |\det D\Phi_s(x)| \,\mathrm{d}\sigma(x)\,\mathrm{d}s \leq Ct^\nu \|h\|_{C^{0,\nu}(\overline{\Gamma_\varepsilon})}$$

and

$$\int_0^t \int_{\partial D} h(x-sn(x))(|\det D\Phi_s(x)| - |\det D\Phi_{-s}(x)|)\,\mathrm{d}\sigma(x)\,\mathrm{d}s \leq Ct\|h\|_{C^0(\overline{\Gamma_\varepsilon})}.$$

The proof is completed by integration over t with similar arguments as in the proof of Theorem 5.2. □

5.2 Convergence of diffuse boundary integrals

In this section we investigate the accuracy of the diffuse boundary integral approximation. For this sake consider

$$E_B = \int_\Omega g(x)|\nabla\omega^\varepsilon(x)|\,\mathrm{d}x - \int_{\partial D} g(x)\,\mathrm{d}\sigma(x).$$

In the following we reduce the treatment of E_B to that of E_V from the previous section. Using $\nabla d_D(x) = n(x)$ and the divergence theorem, we see that

$$\int_{\partial D} g(x)\,\mathrm{d}\sigma(x) = \int_{\partial D} g(x)\nabla d_D(x)\cdot n(x)\,\mathrm{d}\sigma(x) = \int_D \operatorname{div}(g(x)\nabla d_D(x))\,\mathrm{d}x.$$

Note that, $g_{|D} \in W^{1,p}(D)$ for any $g \in W^{1,p}(D_\varepsilon;\omega^\varepsilon)$, and thus g has a trace on ∂D. To treat the diffuse boundary integral, we first observe that $|\nabla\omega^\varepsilon| = -\nabla d_D\cdot\nabla\omega^\varepsilon$ on Γ_ε. For the definition of Γ_ε see (7). Therefore, integration-by-parts shows that

$$\begin{aligned}\int_\Omega g(x)|\nabla\omega^\varepsilon(x)|\,\mathrm{d}x &= -\int_\Omega g(x)\nabla d_D(x)\nabla\omega^\varepsilon(x)\,\mathrm{d}x \\ &= \int_\Omega \operatorname{div}(g(x)\nabla d_D(x))\omega^\varepsilon(x)\,\mathrm{d}x.\end{aligned}$$

Notice, that due to $\operatorname{supp}(\omega^\varepsilon)\subset\Omega$ there are no boundary integrals. Thus, we have that

$$E_B = \int_\Omega \operatorname{div}(g(x)\nabla d_D(x))\,\mathrm{d}\omega^\varepsilon(x) - \int_D \operatorname{div}(g(x)\nabla d_D(x))\,\mathrm{d}x.$$

Setting $h = \operatorname{div}(g\nabla d_D)$, we can use Theorem 5.1, Theorem 5.2 and Lemma 5.3 of the previous section.

Lemma 5.4. *Let ∂D be of class $C^{1,1}$ and let $1 \leq p \leq \infty$. Moreover, let $g \in W^{1,p}(D_\varepsilon;\omega^\varepsilon)$ for some $0 < \varepsilon < \varepsilon_0$. Then there exists a constant $C > 0$ independent of ε such that*

$$|E_B| \leq C\|g\|_{W^{1,p}(D_\varepsilon;\omega^\varepsilon)}\varepsilon^{1-\frac{1}{p}}.$$

Proof. If $\partial D \in C^{1,1}$, then $d_D \in C^{1,1}$ [11] and, in this case, $g \in W^{1,p}(D_\varepsilon;\omega^\varepsilon)$ implies $h \in L^p(D_\varepsilon;\omega^\varepsilon)$ for $1 \leq p \leq \infty$, which in turn implies $E_B = O(\varepsilon^{1-1/p})$ by Theorem 5.1. □

Lemma 5.5. *Let ∂D be of class $C^{2,1}$. Moreover, let $g \in W^{2,p}(D_\varepsilon;\omega^\varepsilon)$ for some $0 < \varepsilon < \varepsilon_0$ and $1 \leq p \leq \infty$. Then there exists a constant $C > 0$ independent of ε such that*

$$|E_B| \leq C\|g\|_{W^{2,p}(D_\varepsilon;\omega^\varepsilon)}\varepsilon^{2-\frac{1}{p}}.$$

Proof. If $\partial D \in C^{2,1}$, then $d_D \in C^{2,1}$ [11] and, in this case, $g \in W^{2,p}(D_\varepsilon;\omega^\varepsilon)$ implies $h \in W^{1,p}(D_\varepsilon;\omega^\varepsilon)$, which in turn implies $E_B = O(\varepsilon^{2-1/p})$ by Theorem 5.2. □

The estimate of Lemma 5.5 assumes $W^{2,p}$-regularity of the whole integrand. For our analysis we will also need a slightly different statement:

Theorem 5.6. *Assume ∂D is of class $C^{1,1}$, let (S3) hold and let $1 \leq p \leq q \leq \infty$. Furthermore, let $u \in W^{2,q}(D_\varepsilon;\omega^\varepsilon)$ satisfy $u = 0$ on ∂D and let $v \in W^{1,p'}(D_\varepsilon;\omega^\varepsilon)$ with $p' = p/(p-1)$. Then there exists a constant C independent of ε, u and v such that for $q' = q/(q-1)$*

$$\int_{\Gamma_\varepsilon} uv|\nabla\omega^\varepsilon|\,\mathrm{d}x \leq C(\varepsilon^{1+\frac{1}{q'}}\|u\|_{W^{2,q}(D_\varepsilon;\omega^\varepsilon)} + \varepsilon^{1+\frac{1}{p}}\|u\|_{W^{2,p}(D_\varepsilon;\omega^\varepsilon)})\|v\|_{W^{1,p'}(D_\varepsilon;\omega^\varepsilon)}.$$

The higher integrability of u improves the first part of the estimate whereas the higher integrability of v improves the second part. For $q = p$ the rate is $O(\varepsilon^{1+\frac{1}{p'}} + \varepsilon^{1+\frac{1}{p}})$ which is optimal for $p = 2$. For $q = p'$ we obtain the best possible rate $O(\varepsilon^{1+\frac{1}{p}})$.

Proof. We start with an inequality for $w \in W^{1,1}(D_\varepsilon;\omega^\varepsilon)$. An application of (11), (9) and Theorem 4.2 yields

$$\begin{aligned}
&\Big|\int_{\Gamma_\varepsilon} w|\nabla\omega^\varepsilon|\,\mathrm{d}x - \int_{-\varepsilon}^{\varepsilon}\frac{1}{2\varepsilon}S'(-\frac{t}{\varepsilon})\int_{\partial D} w(x+tn(x))(1+t\Delta d_D(x))\,\mathrm{d}\sigma(x)\,\mathrm{d}t\Big| \qquad (24)\\
&\qquad\leq C\int_{-\varepsilon}^{\varepsilon}\frac{1}{2\varepsilon}S'(-\frac{t}{\varepsilon})\int_{\partial D}|w(x+tn(x))|\varepsilon^2\,\mathrm{d}\sigma(x)\,\mathrm{d}t\\
&\qquad\leq C\varepsilon^2\int_{\Gamma_\varepsilon}|w||\nabla\omega^\varepsilon|\,\mathrm{d}x\\
&\qquad\leq C\varepsilon^2\|w\|_{W^{1,1}(D_\varepsilon;\omega^\varepsilon)},
\end{aligned}$$

with C independent of ε.
Now let $w \in W^{2,q}((-\varepsilon,\varepsilon))$, then there exists a constant C such that

$$|\int_{-\varepsilon}^{\varepsilon} S'(-\frac{t}{\varepsilon}) \int_0^t w'(s)\,\mathrm{d}s\,\mathrm{d}t| \leq \varepsilon^{3-\frac{3}{q}} C \left(\int_0^{\varepsilon} S'(-\frac{t}{\varepsilon}) \int_0^t \int_{-s}^s |w''(r)|^q\,\mathrm{d}r\,\mathrm{d}s\,\mathrm{d}t \right)^{\frac{1}{q}}. \tag{25}$$

This can be seen as follows: By change of variables and application of the fundamental theorem of calculus, we obtain

$$\begin{aligned}\int_{-\varepsilon}^{\varepsilon} S'(-\frac{t}{\varepsilon}) \int_0^t w'(s)\,\mathrm{d}s\,\mathrm{d}t &= \int_0^{\varepsilon} S'(-\frac{t}{\varepsilon}) \int_0^t w'(s) - w'(-s)\,\mathrm{d}s\,\mathrm{d}t \\ &= \int_0^{\varepsilon} S'(-\frac{t}{\varepsilon}) \int_0^t \int_{-s}^s w''(r)\,\mathrm{d}r\,\mathrm{d}s\,\mathrm{d}t.\end{aligned}$$

Repeated application of Hölder's inequality gives

$$\begin{aligned}&|\int_0^{\varepsilon} S'(-\frac{t}{\varepsilon}) \int_0^t \int_{-s}^s w''(r)\,\mathrm{d}r\,\mathrm{d}s\,\mathrm{d}t| \\ &\leq 2^{\frac{1}{q'}} \left(\int_0^{\varepsilon} S'(-\frac{t}{\varepsilon}) \int_0^t s\,\mathrm{d}s\,\mathrm{d}t \right)^{\frac{1}{q'}} \left(\int_0^{\varepsilon} S'(-\frac{t}{\varepsilon}) \int_0^t \int_{-s}^s |w''(r)|^q\,\mathrm{d}r\,\mathrm{d}s\,\mathrm{d}t \right)^{\frac{1}{q}}.\end{aligned}$$

Using boundedness of S' and calculating $\int_0^{\varepsilon}\int_0^t s\,\mathrm{d}s\,\mathrm{d}t = \varepsilon^3/6$ yields the assertion.
We are now in the position to give a proof of the theorem. By setting $w = uv$ in (24), we have that

$$\begin{aligned}&\int_{\Gamma_\varepsilon} uv|\nabla\omega^{\varepsilon}|\,\mathrm{d}x \\ &- \int_{-\varepsilon}^{\varepsilon} \frac{1}{2\varepsilon} S'(-\frac{t}{\varepsilon}) \int_{\partial D} u(x+tn(x))v(x+tn(x))(1+t\Delta d_D(x))\,\mathrm{d}\sigma(x)\,\mathrm{d}t \\ &\leq C\varepsilon^2 \|u\|_{W^{1,p}(D_\varepsilon;\omega^{\varepsilon})} \|v\|_{W^{1,p'}(D_\varepsilon;\omega^{\varepsilon})}.\end{aligned}$$

Thus, to prove the theorem, it is sufficient to estimate the second integral on the left-hand side of the latter inequality. Using $v(x+tn(x)) = v(x) + \int_0^t \nabla v(x+sn(x))\cdot n(x)\,\mathrm{d}s$ and $u(x+tn(x)) = \int_0^t \nabla u(x+sn(x))\cdot n(x)\,\mathrm{d}s$, we see that

$$\begin{aligned}u(x+tn(x))v(x+tn(x)) &= v(x) \int_0^t \nabla u(x+sn(x))\cdot n(x)\,\mathrm{d}s \\ &\quad + \int_0^t \nabla v(x+sn(x))\cdot n(x)\,\mathrm{d}s \int_0^t \nabla u(x+sn(x))\cdot n(x)\,\mathrm{d}s.\end{aligned}$$

We treat the two terms on the right-hand side separately. For the first one, we will use (25) with $w(s) = u(x+sn(x))$, Hölder's inequality and $q \geq p$

which yields

$$
\begin{aligned}
&\left|\frac{1}{2\varepsilon}\int_{\partial D} v(x)\int_{-\varepsilon}^{\varepsilon} S'(-\frac{t}{\varepsilon})\int_0^t \nabla u(x+sn(x))\cdot n(x)\,\mathrm{d}s\,\mathrm{d}t\,\mathrm{d}\sigma(x)\right| \\
&\le C\frac{\varepsilon^{3-\frac{3}{q}}}{2\varepsilon}\int_{\partial D} v(x)\left(\int_0^{\varepsilon} S'(-\frac{t}{\varepsilon})\right. \\
&\qquad\left.\int_0^t\int_{-s}^{s} |n(x)\cdot D^2u(x+rn(x))\cdot n(x)|^q dr\,\mathrm{d}s\,\mathrm{d}t\right)^{\frac{1}{q}}\mathrm{d}\sigma(x) \\
&\le C\varepsilon^{2-\frac{3}{q}}\left(\int_{\partial D}|v|^{p'}\,\mathrm{d}\sigma\right)^{\frac{1}{p'}}\left(\int_0^{\varepsilon} S'(-\frac{t}{\varepsilon})\int_0^t\int_{\Gamma_s}|D^2u|^q\,\mathrm{d}x\,\mathrm{d}s\,\mathrm{d}t\right)^{\frac{1}{q}} \\
&\le C\varepsilon^{2-\frac{1}{q}}\left(\int_{\partial D}|v|^{p'}\,\mathrm{d}\sigma\right)^{\frac{1}{p'}}\left(\int_0^{\varepsilon}\frac{1}{2\varepsilon}S'(-\frac{t}{\varepsilon})\int_{\Gamma_t}|D^2u|^q\,\mathrm{d}x\,\mathrm{d}t\right)^{\frac{1}{q}} \\
&\le C\varepsilon^{2-\frac{1}{q}}\|v\|_{W^{1,p'}(D_\varepsilon;\omega^\varepsilon)}\|u\|_{W^{2,q}(D_\varepsilon;\omega^\varepsilon)},
\end{aligned}
$$

where we have used Lemma 4.1 to treat the term involving v and the transformation formula to treat the term involving u, see the last lines of the proof of Theorem 5.2. For the second term we first use Hölder's inequality twice

$$
\begin{aligned}
&\Big|\frac{1}{2\varepsilon}\int_{\partial D}\int_0^{\varepsilon} S'(-\frac{t}{\varepsilon})\left(\int_0^t \nabla v(x+sn(x))\cdot n(x)\,\mathrm{d}s\right. \\
&\qquad\left.\int_0^t \nabla u(x+sn(x))\cdot n(x)\,\mathrm{d}s\right)\,\mathrm{d}t\,\mathrm{d}\sigma(x)\Big| \\
&\le \frac{1}{2\varepsilon}\int_{\partial D}\int_0^{\varepsilon} tS'(-\frac{t}{\varepsilon})\left(\int_0^t|\nabla v(x+sn(x))|^{p'}\,\mathrm{d}s\right)^{\frac{1}{p'}} \\
&\qquad\qquad\left(\int_0^t|\nabla u(x+sn(x))|^{p}\,\mathrm{d}s\right)^{\frac{1}{p}}\,\mathrm{d}t\,\mathrm{d}\sigma(x) \\
&\le \frac{\varepsilon}{2}\left(\int_{\partial D}\frac{1}{\varepsilon}\int_0^{\varepsilon} S'(-\frac{t}{\varepsilon})\int_0^t|\nabla v(x+sn(x))|^{p'}\,\mathrm{d}s\,\mathrm{d}t\,\mathrm{d}\sigma(x)\right)^{\frac{1}{p'}} \\
&\qquad\left(\int_{\partial D}\frac{1}{\varepsilon}\int_0^{\varepsilon} S'(-\frac{t}{\varepsilon})\int_0^t|\nabla u(x+sn(x))|^{p}\,\mathrm{d}s\,\mathrm{d}t\,\mathrm{d}\sigma(x)\right)^{\frac{1}{p}}.
\end{aligned}
$$

Then, using (11), we obtain similarly as in the proof of Theorem 5.2

$$
\int_{\partial D}\frac{1}{\varepsilon}\int_0^{\varepsilon} S'(-\frac{t}{\varepsilon})\int_0^t|\nabla v(x+sn(x))|^{p'}\,\mathrm{d}s\,\mathrm{d}t\,\mathrm{d}\sigma(x)\le C\|v\|^{p'}_{W^{1,p'}(D_\varepsilon;\omega^\varepsilon)},
$$

and by (11), (S3), and by Theorem 4.2

$$\begin{aligned}
&\int_{\partial D} \frac{1}{\varepsilon} \int_0^\varepsilon S'(-\frac{t}{\varepsilon}) \int_0^t |\nabla u(x+sn(x))|^p \,\mathrm{d}s \,\mathrm{d}t \,\mathrm{d}\sigma(x) \\
&\leq \varepsilon \int_{\partial D} \int_0^\varepsilon \frac{1}{\varepsilon} S'(-\frac{s}{\varepsilon}) |\nabla u(x+sn(x))|^p \,\mathrm{d}s \,\mathrm{d}\sigma(x) \\
&\leq C\varepsilon \int_{\Gamma_\varepsilon} |\nabla u|^p |\nabla \omega^\varepsilon| \,\mathrm{d}x \\
&\leq C\varepsilon \|u\|^p_{W^{2,p}(D_\varepsilon;\omega^\varepsilon)}.
\end{aligned}$$

The integrals over $(-\varepsilon, 0)$ as well as the ones involving $t\Delta d_D$ can be treated similarly. Collecting all terms yields the assertion. □

Up to now, we have always assumed the boundary data g to be regular. For completeness, let us also consider the case $g \in L^p(\partial D)$ only. Then g is defined a.e. on ∂D, and we can define an extension a.e. on Γ_ε by

$$\tilde{g}(x + tn(x)) = g(x), \qquad -\varepsilon \leq t \leq \varepsilon,\ x \in \partial D. \tag{26}$$

Lemma 5.7. *Let $g \in L^p(\partial D)$ and let $v \in W^{1,p'}(D_\varepsilon;\omega^\varepsilon)$ with $1 \leq p \leq \infty$, $p' = p/(p-1)$, and $0 < \varepsilon \leq \varepsilon_0$. Then there exists a constant C independent of ε such that*

$$\Big| \int_{\Gamma_\varepsilon} \tilde{g} v |\nabla \omega^\varepsilon| \,\mathrm{d}x - \int_{\partial D} gv \,\mathrm{d}\sigma \Big| \leq C\varepsilon^{1/p} \|g\|_{L^p(\partial D)} \|v\|_{W^{1,p'}(D_\varepsilon;\omega^\varepsilon)}$$

with $\tilde{g}$ being the extension defined in (26).

Proof. Using $\int_{-\varepsilon}^{\varepsilon} S'(-t/\varepsilon)\,\mathrm{d}t = 2\varepsilon$ and the transformation formula, we obtain

$$\begin{aligned}
&\int_{\Gamma_\varepsilon} \tilde{g} v |\nabla \omega^\varepsilon| \,\mathrm{d}x - \int_{\partial D} gv \,\mathrm{d}\sigma \\
&= \int_{-\varepsilon}^{\varepsilon} \frac{1}{2\varepsilon} S'(-\frac{t}{\varepsilon}) \int_{\partial D} \big(v(x+tn(x)) \det D\Phi_t(x) - v(x)\big) g(x) \,\mathrm{d}\sigma(x) \,\mathrm{d}t.
\end{aligned}$$

Thus, using (9), there exists $C > 0$ independent of ε such that

$$\begin{aligned}
&\Big| \int_{\Gamma_\varepsilon} \tilde{g} v |\nabla \omega^\varepsilon| \,\mathrm{d}x - \int_{\partial D} gv \,\mathrm{d}\sigma \Big| \\
&\qquad \leq \left| \int_{\partial D} g(x) \int_{-\varepsilon}^{\varepsilon} \frac{1}{2\varepsilon} S'(-\frac{t}{\varepsilon}) \big(v(x+tn(x)) - v(x)\big) \,\mathrm{d}t \,\mathrm{d}\sigma(x) \right| \\
&\qquad + C \left| \int_{\partial D} g(x) \int_{-\varepsilon}^{\varepsilon} \frac{1}{2\varepsilon} S'(-\frac{t}{\varepsilon}) t v(x+tn(x)) \,\mathrm{d}t \,\mathrm{d}\sigma(x) \right|.
\end{aligned}$$

We treat the two integrals on the right-hand side separately. Repeated use of Hölder's inequality and $\nabla v(x+tn(x)) - v(x) = \int_0^t \nabla v(x+sn(x)) \cdot n(x)\,\mathrm{d}s$ yields similarly as in the proof of Theorem 5.2

$$
\begin{aligned}
&\Big| \int_{\partial D} g(x) \int_0^\varepsilon \frac{1}{2\varepsilon} S'(-\frac{t}{\varepsilon}) \big(v(x+tn(x)) - v(x)\big)\,\mathrm{d}t\,\mathrm{d}\sigma(x)\Big| \\
&\leq \|g\|_{L^p(\partial D)} \left(\int_{\partial D} \left(\int_0^\varepsilon \frac{1}{2\varepsilon} S'(-\frac{t}{\varepsilon}) \int_0^t |\nabla v(x+sn(x))| ds\,\mathrm{d}t \right)^{p'} \mathrm{d}\sigma(x) \right)^{\frac{1}{p'}} \\
&\leq \|g\|_{L^p(\partial D)} |\Gamma_\varepsilon|^{\frac{1}{p}} \left(\int_0^\varepsilon \frac{1}{2\varepsilon} S'(-\frac{t}{\varepsilon}) \int_{\Gamma_t} |\nabla v(x)|^{p'}\,\mathrm{d}x\,\mathrm{d}t \right)^{p'}.
\end{aligned}
$$

An analogue estimate hold for the integral over $(-\varepsilon, 0)$. Since, $|\Gamma_\varepsilon|^{\frac{1}{p}} \leq C\varepsilon^{1/p}$ this is the desired estimate for the first integral. The second can be estimated similarly, i.e.

$$
\begin{aligned}
&\left| \int_{\partial D} g(x) \int_{-\varepsilon}^\varepsilon \frac{1}{2\varepsilon} S'(-\frac{t}{\varepsilon}) t v(x+tn(x))\,\mathrm{d}t\,\mathrm{d}\sigma(x) \right| \\
&\qquad \leq \varepsilon \|g\|_{L^p(\partial D)} \left(\int_{\Gamma_\varepsilon} |v|^{p'} |\nabla \omega^\varepsilon|\,\mathrm{d}x \right)^{\frac{1}{p'}}
\end{aligned}
$$

The last term can be estimated using the trace theorem 4.2. □

For the sake of completeness, we also state an analog to Lemma 5.3.

Lemma 5.8. *Let ∂D be of class $C^{2,\nu}$ for some $0 < \nu \leq 1$. Moreover, let $g \in C^{1,\nu}(\overline{\Gamma_\varepsilon})$ for some $0 < \varepsilon < \varepsilon_0$. Then there exists a constant $C > 0$ independent of ε such that*

$$
|E_B| \leq C \|g\|_{C^{1,\nu}(\overline{\Gamma_\varepsilon})} \varepsilon^{1+\nu}.
$$

Proof. Since $d_D \in C^{2,\nu}(\overline{\Gamma_\varepsilon})$, we have $g \nabla d_D \in C^{1,\nu}(\overline{\Gamma_\varepsilon})$. The assertion follows from Lemma 5.3. □

6 Diffuse elliptic problems

In this section we investigate three typical second order elliptic boundary value problems. We start with Robin-type problems, which build the basis for further investigations. For rather irregular data, we obtain a weak sublinear convergence result in terms of ε. Superlinear convergence is achieved by requiring smooth data. In Section 6.2 we treat Dirichlet boundary conditions which can be reduced to the analysis of a Robin problem by means of the well-known penalty method. In Section 6.3 we consider Neumann boundary conditions and establish well-posedness of the diffuse domain method. Apart from the well-posed the convergence results can be derived as in the Robin case.

6.1 Robin boundary conditions

Consider the following second order elliptic equation with Robin-type boundary condition: Find u such that

$$-\mathrm{div}(A\nabla u) + cu = f \quad \text{in } D, \tag{27}$$
$$n \cdot A\nabla u + bu = g \quad \text{on } \partial D. \tag{28}$$

In order to obtain (weak) solutions to (27)–(28), let us consider the following weak formulation: Find $u \in W^{1,2}(D)$ such that

$$a(u,v) = \ell(v) \quad \text{for all } v \in W^{1,2}(D), \tag{29}$$

with bilinear and linear form

$$a(u,v) = \int_D A\nabla u \cdot \nabla v + cuv \,\mathrm{d}x + \int_{\partial D} buv \,\mathrm{d}\sigma,$$
$$\ell(v) = \int_D fv \,\mathrm{d}x + \int_{\partial D} gv \,\mathrm{d}\sigma.$$

In order to prove well-posedness of the weak form (29) via the Lax-Milgram lemma we make the following assumptions:

(C1) $0 < b_0 \leq b \in W^{1,\infty}(\Omega)$, $0 \leq c \in L^\infty(\Omega)$.

(C2) $A \in L^\infty(\Omega)^{n\times n}$ is a symmetric positive definite matrix, i.e. there exists $\kappa > 0$ such that for a.e. $x \in \Omega$

$$\kappa^{-1}|\xi|^2 \leq \xi \cdot A(x)\xi \leq \kappa|\xi|^2 \quad \text{for all } \xi \in \mathbb{R}^n.$$

Lemma 6.1. *Let (C1)–(C2) hold. Moreover, let $f \in L^2(D)$ and $g \in W^{1,2}(D)$. Then there exists a unique $u \in W^{1,2}(D)$ satisfying* (29)*, and there exists $C > 0$ such that*

$$\|u\|_{W^{1,2}(D)} \leq C(\|f\|_{L^2(D)} + \|g\|_{L^2(\partial D)}).$$

The diffuse approximation of (29) is now: Find $u^\varepsilon \in W^{1,2}(D_\varepsilon;\omega^\varepsilon)$ such that

$$a^\varepsilon(u^\varepsilon, v) = \ell^\varepsilon(v) \quad \text{for all } v \in W^{1,2}(D_\varepsilon;\omega^\varepsilon), \tag{30}$$

where the corresponding bilinear and linear form are given by

$$a^\varepsilon(u^\varepsilon, v) = \int_\Omega A\nabla u^\varepsilon \cdot \nabla v + cu^\varepsilon v \,\mathrm{d}\omega^\varepsilon + \int_\Omega bu^\varepsilon v|\nabla\omega^\varepsilon| \,\mathrm{d}x$$
$$\ell^\varepsilon(v) = \int_\Omega fv \,\mathrm{d}\omega^\varepsilon + \int_\Omega gv|\nabla\omega^\varepsilon| \,\mathrm{d}x.$$

Lemma 6.2. *Let (C1)–(C2) hold. Moreover, let $f \in L^2(D_\varepsilon;\omega^\varepsilon)$ and $g \in W^{1,2}(D_\varepsilon;\omega^\varepsilon)$. Then there exists a unique $u^\varepsilon \in W^{1,2}(D_\varepsilon;\omega^\varepsilon)$ satisfying* (30)*, and there exists $C>0$ independent of ε such that*

$$\|u^\varepsilon\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)} \le C(\|f\|_{L^2(D_\varepsilon;\omega^\varepsilon)} + \|g\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)}).$$

Proof. Continuity of a^ε and ℓ^ε with respect to the $W^{1,2}(D_\varepsilon;\omega^\varepsilon)$-topology follows from boundedness of the coefficients and Theorem 4.2. Coercivity of a^ε on $W^{1,2}(D_\varepsilon;\omega^\varepsilon)$ is a direct consequence of the positivity of A and the Poincaré-Friedrichs inequality, see Corollary 4.11. An application of the Lax-Milgram lemma yields the assertion. □

Denoting by u and u^ε the corresponding solutions to (29) and (30), respectively, we next want to estimate the error $u-u^\varepsilon$ with respect to the $W^{1,2}(D_\varepsilon;\omega^\varepsilon)$-norm which directly implies estimates in the $W^{1,2}(D)$-norm as well. By regularity of ∂D, we can assume that $u : D \to \mathbb{R}$ is extended to Ω preserving $W^{1,2}(\Omega)$-regularity. Hence, the error $u-u^\varepsilon$ satisfies

$$a^\varepsilon(u-u^\varepsilon,v) = a^\varepsilon(u,v) - a(u,v) + \ell(v) - \ell^\varepsilon(v) \quad \text{for all } v \in W^{1,2}(D_\varepsilon;\omega^\varepsilon). \tag{31}$$

6.1.1 Sublinear convergence

In order to obtain a first estimate for the error $u-u^\varepsilon$, we estimate the right-hand side of (31) by employing the embedding theorem 4.3. We recall the definitions $p^*_\alpha = (n+\alpha)p/(n+\alpha-p)$, see (14), and

$$\|\ell\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)'} = \sup_{v\in W^{1,2}(D_\varepsilon;\omega^\varepsilon)} \frac{\ell(v)}{\|v\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)}},$$

which is the norm of ℓ as an element of the dual space of $W^{1,2}(D_\varepsilon;\omega^\varepsilon)$.

Lemma 6.3. *Let $f \in L^2(D_\varepsilon,\omega^\varepsilon)$ and $g \in W^{1,2}(D_\varepsilon;\omega^\varepsilon)$. Then there exists a constant C independent of ε such that*

$$\|\ell^\varepsilon - \ell\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)'} \le C\left(\|f\|_{L^2(D_\varepsilon;\omega^\varepsilon)} + \|g\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)}\right)\varepsilon^{\frac{1}{n+\alpha}}.$$

Proof. Let $v \in W^{1,2}(D_\varepsilon;\omega^\varepsilon)$. Due to the weighted Sobolev embedding (15) we have $v \in L^p(D_\varepsilon,\omega^\varepsilon)$ for $p = 2^*_\alpha$. Hence, $fv \in L^q(D_\varepsilon,\omega^\varepsilon)$ with $q = 2p/(2+p)$ due to Hölder's inequality. Similarly, since $\nabla(gv) = g\nabla v + \nabla g v \in L^q(D_\varepsilon,\omega^\varepsilon)$ with q as before, we have that $gv \in W^{1,q}(D_\varepsilon,\omega^\varepsilon)$. Using Theorem 5.1 and Lemma 5.4 we obtain

$$|\ell^\varepsilon(v) - \ell(v)| \le C\left(\|f\|_{L^2(D_\varepsilon;\omega^\varepsilon)} + \|g\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)}\right)\|v\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)}\varepsilon^{1-\frac{1}{q}}$$

The assertion follows from $1-\frac{1}{q} = \frac{1}{2} - \frac{1}{2^*_\alpha} = \frac{1}{n+\alpha}$. □

In order to obtain convergence rates, we need some regularity of u.

Lemma 6.4. *Let $u \in W^{1,p}(D_\varepsilon;\omega^\varepsilon)$ for some $p > 2$. Then there exists a constant C independent of ε such that*

$$\|a^\varepsilon(u,\cdot) - a(u,\cdot)\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)'} \le C\|u\|_{W^{1,p}(D_\varepsilon;\omega^\varepsilon)}\varepsilon^{\frac{1}{2}-\frac{1}{p}}.$$

Proof. Let $v \in W^{1,2}(D_\varepsilon;\omega^\varepsilon)$ be arbitrary. Since A is bounded and $u \in W^{1,p}(D_\varepsilon;\omega^\varepsilon)$, it is $A\nabla u \cdot \nabla v \in L^q(D_\varepsilon;\omega^\varepsilon)$ with $q = 2p/(2+p)$. Similarly $cuv \in L^q(D_\varepsilon;\omega^\varepsilon)$. Using $\alpha \in W^{1,\infty}(\Omega)$, we see that $\alpha uv \in W^{1,q}(D_\varepsilon;\omega^\varepsilon)$ with q as before. The result now follows by applying Theorem 5.1 and Lemma 5.4 similar as in the proof of Lemma 6.3. □

Having estimated the errors in right-hand side and bilinear form we can proceed to the main approximation results in this section:

Theorem 6.5. *Let (C1)–(C2) hold. Moreover, assume that $u \in W^{1,p}(D)$ with $2 \le p \le 2^*_\alpha$ is a solution to (29) and $u^\varepsilon \in W^{1,2}(D_\varepsilon;\omega^\varepsilon)$ is a solution to (30). Then there exists a constant $C > 0$ independent of ε such that*

$$\begin{aligned}&\|u - u^\varepsilon\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)}\\ &\qquad \le C\left(\|u\|_{W^{1,p}(D_\varepsilon;\omega^\varepsilon)} + \|f\|_{L^2(D_\varepsilon;\omega^\varepsilon)} + \|g\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)}\right)\varepsilon^{\frac{1}{2}-\frac{1}{p}}.\end{aligned}$$

Proof. Coercivity of a^ε and (31) imply

$$\begin{aligned}&\|u - u^\varepsilon\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)}\\ &\qquad \le C(\|a^\varepsilon(u,\cdot) - a(u,\cdot)\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)'} + \|\ell^\varepsilon - \ell\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)'}),\end{aligned}$$

and the assertion follows from Lemma 6.3 and Lemma 6.4. □

Remark 6.6. *Note that, according to [18], see also [12], there always exists a $p > 2$ such that $u \in W^{1,p}(D)$, whence $u \in W^{1,p}(D_\varepsilon;\omega^\varepsilon)$ by extension. If $p = 2^*_\alpha$, we obtain the best possible rate $O(\varepsilon^{1/(n+\alpha)})$, i.e. $O(\varepsilon^{1/3})$ in two space dimensions and S as in Example 3.1 (i).*

Remark 6.7. *(i) Assuming $g = 0$ and $f \in L^2(D)$ extended by zero to Ω an inspection of the proof of Theorem 5.1 shows that for each $v \in W^{1,2}(D_\varepsilon;\omega^\varepsilon)$*

$$\int_{D_\varepsilon} fv \,\mathrm{d}\omega^\varepsilon - \int_D fv \,\mathrm{d}x \le C\varepsilon^{\frac{1}{n}}\|v\|_{W^{1,2}(D)}\|f\|_{L^2(D)}$$

which is due to the embedding $W^{1,2}(D) \hookrightarrow L^{2^*_0}(D)$ and the fact that $v_{|D} \in W^{1,2}(D)$. This immediately leads to a stronger result in Lemma 6.3 independent of ω^ε. However, for proving Lemma 6.4, we have to estimate the term

$$\int_{D_\varepsilon} A\nabla u \cdot \nabla v \,\mathrm{d}\omega^\varepsilon - \int_D A\nabla u \cdot \nabla v \,\mathrm{d}x.$$

Here, on the one hand, to preserve regularity of u, setting $u = 0$ on $D_\varepsilon \setminus D$ is not possible. On the other hand setting $A = 0$ on $D_\varepsilon \setminus D$ is not allowed since then (6) is not well-posed anymore.
(ii) If $f \in L^\infty(\Omega)$, $g \in W^{1,\infty}(\Omega)$ and $u \in W^{1,\infty}(\Omega)$, then using the techniques from above, we would obtain the bound

$$\|u - u^\varepsilon\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)} = O(\varepsilon^{\frac{1}{2}})$$

as $\varepsilon \to 0$ since the test function v is merely $W^{1,2}(D_\varepsilon;\omega^\varepsilon)$. For smooth test functions v and $f \in L^p(D_\varepsilon;\omega^\varepsilon)$ and $u, g \in W^{1,p}(D_\varepsilon;\omega^\varepsilon)$, the right-hand side of (31) is bounded by a constant multiple (depending on f, g and u) of $\varepsilon^{1-1/p}\|v\|_{W^{1,\infty}(D_\varepsilon)}$. This estimate, however, does not lead to $W^{1,2}(D_\varepsilon;\omega^\varepsilon)$-estimates for the error anymore. We will return to these type of estimates in the next section. We also mention that an inspection of several proofs above shows that crucial terms drop out if the involved functions are symmetric with respect to ∂D (mirrored along the normal direction). Thus, using symmetric extensions of data and solutions as well as a restriction to a Sobolev space of functions symmetric with respect to ∂D could give higher order rates. However, since this does not correspond to the computational practice and would extremely complicate the numerical solution, this seems not of particular practical relevance and hence we do not pursue this direction further.

6.1.2 From linear to quadratic convergence

In literature there exist very recent formal results for the diffuse domain method that state a rate of convergence for the L^2-norm of $O(\varepsilon^2)$ for the Poisson equation with Robin boundary conditions [22]. To give a precise and rigorous statement of such a better rate we need additional regularity of the domain, the data and the solutions. Furthermore, we resort also to other functions spaces. For a smooth function v and $1 \le p \le \infty$ we let $p' = p/(p-1)$ and define

$$\begin{aligned}\|v\|_{\mathcal{X}_p^\varepsilon} &= \|a^\varepsilon(v,\cdot)\|_{W^{1,p'}(D_\varepsilon;\omega^\varepsilon)'} \\ &= \sup\{a^\varepsilon(v,\phi) :\ \phi \in C^\infty(\overline{D_\varepsilon}), \|\phi\|_{W^{1,p'}(D_\varepsilon;\omega^\varepsilon)} \le 1\}.\end{aligned}$$

We let $\mathcal{X}_p^\varepsilon = \{v \in C^\infty(\overline{D_\varepsilon}) :\ \|v\|_{\mathcal{X}_p^\varepsilon} < \infty\}$ denote the completion of $C^\infty(\overline{D_\varepsilon})$ with respect to $\|\cdot\|_{\mathcal{X}_p^\varepsilon}$. Then $(\mathcal{X}_p^\varepsilon, \|\cdot\|_{\mathcal{X}_p^\varepsilon})$ is a Banach space. Due to the Riesz representation theorem, Corollary 4.11, and assumptions (C1)-(C2) on the coefficients, we see that $\mathcal{X}_2^\varepsilon = W^{1,2}(D_\varepsilon;\omega^\varepsilon)$ with equivalent norms. Furthermore, due to Theorem 4.2 we easily see that $\|u\|_{\mathcal{X}_p^\varepsilon} \le C\|u\|_{W^{1,p}(D_\varepsilon;\omega^\varepsilon)}$. For the other direction, we need a solvability result. If for any $\ell \in W^{1,p'}(D_\varepsilon;\omega^\varepsilon)'$ there exists $u \in W^{1,p}(D_\varepsilon;\omega^\varepsilon)$ such that $a^\varepsilon(u,v) = \ell(v)$ for all $v \in W^{1,p'}(D_\varepsilon;\omega^\varepsilon)$ and $\|u\|_{W^{1,p}(D_\varepsilon;\omega^\varepsilon)} \le C\|\ell\|_{W^{1,p'}(D_\varepsilon;\omega^\varepsilon)'}$ for

some constant C, then $\|u\|_{W^{1,p}(D_\varepsilon;\omega^\varepsilon)} \le \|u\|_{\mathcal{X}_p^\varepsilon}$. Let us emphasize that such a result is not known to us for the case $p \neq 2$; we refer to [12, 18] for a corresponding result in the unweighted case.
Let us thus start with an error estimate in $\mathcal{X}_p^\varepsilon$. For simplicity, we will assume smooth data. It should become clear from the proof how to lower these regularity assumptions.

Theorem 6.8. *Let ∂D be of class C^∞, and let $f, g \in C^\infty(\overline{\Omega})$ and let (C1)–(C2) hold. Moreover, let $A \in C^\infty(\overline{\Omega})^{3\times 3}$, $c \in C^\infty(\overline{\Omega})$, $b \in C^\infty(\overline{\Omega})$, and let $u^\varepsilon \in W^{1,2}(D_\varepsilon;\omega^\varepsilon)$ denote the solution to (30), and let $u \in W^{1,2}(D)$ denote the solution to (29). Then, for $1 \le p \le \infty$ there exists a constant C independent of ε such that*

$$\|u - u^\varepsilon\|_{\mathcal{X}_p^\varepsilon} \le C\varepsilon^{1+\frac{1}{p}}.$$

Proof. Due to [17, Thm. 2.4.2.7, Rem. 2.5.1.2] and the smoothness of the data, we have that $u \in W^{k,2}(D)$ for any $k \in \mathbb{N}$, i.e. $u \in C^\infty(\overline{D})$ by embedding, and u is a classical solution to (27)–(28). Therefore, integrating by parts on the right hand side of (31), we deduce that for any $v \in W^{1,p'}(D_\varepsilon;\omega^\varepsilon)$ the error satisfies

$$\begin{aligned} a^\varepsilon(u-u^\varepsilon, v) = &\int_D \operatorname{div}(A\nabla u)v \,\mathrm{d}x - \int_{D_\varepsilon} \operatorname{div}(A\nabla u)v \,\mathrm{d}\omega^\varepsilon + \int_{D_\varepsilon} cuv \,\mathrm{d}\omega^\varepsilon \\ &- \int_D cuv \,\mathrm{d}x - \int_{D_\varepsilon} A\nabla u \cdot \nabla\omega^\varepsilon v \,\mathrm{d}x + \int_{D_\varepsilon} buv|\nabla\omega^\varepsilon| \,\mathrm{d}x \\ &- \int_{D_\varepsilon} gv|\nabla\omega^\varepsilon| \,\mathrm{d}x + \int_D fv \,\mathrm{d}x - \int_{D_\varepsilon} fv \,\mathrm{d}\omega^\varepsilon. \end{aligned}$$

In view of Theorem 5.2 we have the estimates

$$\begin{aligned} \Big|\int_{D_\varepsilon} \operatorname{div}(A\nabla u)v \,\mathrm{d}\omega^\varepsilon - \int_D \operatorname{div}(A\nabla u)v \,\mathrm{d}x\Big| &\\ \le C\varepsilon^{2-\frac{1}{p'}} \|\operatorname{div}(A\nabla u)\|_{W^{1,\infty}(D_\varepsilon)} \|v\|_{W^{1,p'}(D_\varepsilon;\omega^\varepsilon)}&, \\ \Big|\int_{D_\varepsilon} cuv \,\mathrm{d}\omega^\varepsilon - \int_D cuv \,\mathrm{d}x\Big| \le C\varepsilon^{2-\frac{1}{p'}} \|cu\|_{W^{1,\infty}(D_\varepsilon)} \|v\|_{W^{1,p'}(D_\varepsilon;\omega^\varepsilon)}&, \\ \Big|\int_D fv \,\mathrm{d}x - \int_{D_\varepsilon} fv \,\mathrm{d}\omega^\varepsilon\Big| \le C\varepsilon^{2-\frac{1}{p'}} \|f\|_{W^{1,\infty}(D_\varepsilon)} \|v\|_{W^{1,p'}(D_\varepsilon;\omega^\varepsilon)}&. \end{aligned}$$

Since $\nabla\omega^\varepsilon = -n|\nabla\omega^\varepsilon|$, the remaining terms can be estimated as follows

$$\begin{aligned} &\int_{D_\varepsilon} (n \cdot A\nabla u + bu - g)v|\nabla\omega^\varepsilon| \,\mathrm{d}x \\ &\qquad \le C\varepsilon^{1+\frac{1}{p}} \|n \cdot A\nabla u + bu - g\|_{W^{2,\max\{p,p'\}}(D_\varepsilon;\omega^\varepsilon)} \|v\|_{W^{1,p'}(D_\varepsilon;\omega^\varepsilon)} \end{aligned}$$

where we used $n \cdot A\nabla u + bu - g = 0$ on ∂D and Theorem 5.6. Hence, taking the supremum over all $v \in W^{1,p'}(D_\varepsilon;\omega^\varepsilon)$ and observing that $2 - \frac{1}{p'} = 1 + \frac{1}{p}$ yields the assertion. □

Corollary 6.9. *Let the assumptions of Theorem 6.8 hold true. Then, there exists a constant C independent of ε such that*

$$\|u - u^\varepsilon\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)} \leq C\varepsilon^{\frac{3}{2}}.$$

Proof. Set $p = 2$ in Theorem 6.8. The assertion follows from the fact that $\mathcal{X}_2^\varepsilon = W^{1,2}(D_\varepsilon;\omega^\varepsilon)$ with equivalent norms. □

Remark 6.10. *Setting $p = 1$ in Theorem 6.8, we obtain $\|u - u^\varepsilon\|_{\mathcal{X}_1^\varepsilon} \leq C\varepsilon^2$. Let us assume that the norms of $W^{1,1}(D_\varepsilon;\omega^\varepsilon)$ and $\mathcal{X}_1^\varepsilon$ are equivalent (uniform with respect to ε). Then continuity of the embedding $W^{1,1}(D) \hookrightarrow L^{\frac{n}{n-1}}(D)$ implies the existence of a constant C independent of ε such that*

$$\|u - u^\varepsilon\|_{L^{\frac{n}{n-1}}(D)} \leq C\varepsilon^2.$$

In particular for $n = 1$, we obtain $\|u-u^\varepsilon\|_{L^p(D)} = O(\varepsilon^2)$ for any $1 \leq p \leq \infty$, and for $n = 2$ we obtain $\|u - u^\varepsilon\|_{L^2(D)} = O(\varepsilon^2)$, thus we recover the formal results of [22].

6.2 Dirichlet boundary conditions

In this section we consider the diffuse domain approximation of second order elliptic equations with Dirichlet boundary conditions: Find u such that

$$-\mathrm{div}(A\nabla u) + cu = f \quad \text{in } D, \tag{32}$$

$$u = g \quad \text{on } \partial D. \tag{33}$$

In order to obtain (weak) solutions to (32)–(33), let us consider the following weak formulation: Find $u \in W^{1,2}(D)$ such that

$$a(u, v) = \ell(v) \quad \text{for all } v \in W_0^{1,2}(D) \text{ such that } u = g \text{ on } \partial D, \tag{34}$$

with bilinear and linear form

$$a(u, v) = \int_D A\nabla u \cdot \nabla v + cuv \,\mathrm{d}x, \qquad \ell(v) = \int_D fv \,\mathrm{d}x.$$

Here $W_0^{1,2}(D)$ is the kernel of the trace operator on $W^{1,2}(D)$. The weak form (34) is well-posed under assumptions (C1)–(C2) which is shown by using the Lax-Milgram lemma. It is well-known that the solution u to (34) is characterized as the solution of the minimization problem

$$a(v, v) - \ell(v) \to \min_{v \in W^{1,2}(D)} \text{ such that } v = g \text{ in } W^{1/2,2}(\partial D).$$

Using the Lagrange formalism this constrained optimization problem is equivalent to finding a saddle-point $(u, \lambda) \in W^{1,2}(D) \times W^{-1/2,2}(\partial D)$ of the Lagrangian

$$L(v, \mu) = a(v, v) - \ell(v) - \langle \mu, g - v \rangle \quad \text{with } v \in W^{1,2}(D),\ \mu \in W^{-1/2,2}(\partial D). \tag{35}$$

Here, $W^{-1/2,2}(\partial D)$ is the topological dual space of the $W^{1,2}(D)$-trace space $W^{1/2,2}(\partial D)$, and $\langle\cdot,\cdot\rangle$ denotes the duality pairing between $W^{-1/2,2}(\partial D)$ and $W^{1/2,2}(\partial D)$. The variational characterization of the saddle-point problem is the following: Find $(u,\lambda)\in W^{1,2}(D)\times W^{-1/2,2}(\partial D)$ such that

$$a(u,v)+\langle\lambda,v\rangle=\ell(v)\qquad\text{for all } v\in W^{1,2}(D), \tag{36}$$

$$\langle\mu,u\rangle=\langle\mu,g\rangle\qquad\text{for all } \mu\in W^{-1/2,2}(\partial D). \tag{37}$$

We have by definition of the norm on $W^{-1/2,2}(\partial D)$ that

$$\|\mu\|_{W^{-1/2,2}(\partial D)}=\sup_{v\in W^{1/2,2}(\partial D)\setminus\{0\}}\frac{\langle\mu,v\rangle}{\|v\|_{W^{1/2,2}(\partial D)}}$$

which asserts an inf-sup condition for the bilinear form $(\mu,v)\mapsto\langle\mu,v\rangle$. Well-posedness of the latter saddle-point problem can then be shown by using Brezzi's splitting theorem [10], cf. [9, Chapter III]. Next, let us introduce a penalized version of (36)–(37) which establishes a connection to elliptic problems with Robin boundary condition discussed in Section 6.1: Let $\beta>0$. Find $(u_\beta,\lambda_\beta)\in W^{1,2}(D)\times W^{-1/2,2}(\partial D)$ such that

$$a(u_\beta,v)+\langle\lambda_\beta,v\rangle=\ell(v)\qquad\text{for all } v\in W^{1,2}(D), \tag{38}$$

$$\langle\mu,u_\beta\rangle-\beta\langle\lambda_\beta,\mu\rangle=\langle\mu,g\rangle\qquad\text{for all } \mu\in W^{-1/2,2}(\partial D). \tag{39}$$

In slight abuse of notation, $\langle\lambda,\mu\rangle$ denotes the inner product on $W^{-1/2,2}(D)$, and is defined as $\langle\lambda,\mu\rangle_{W^{-1/2,2}(\partial D)}=\langle J\lambda,J\mu\rangle_{W^{1/2,2}(\partial D)}$. Here,

$$J:W^{-1/2,2}(\partial D)\to W^{1/2,2}(\partial D)$$

is the Riesz isomorphism and $J\lambda$ is given as the trace of the solution to the Neumann problem

$$-\Delta w+w=0\quad\text{in } D,\qquad\partial_n w=\lambda\quad\text{on } \partial D.$$

We have that

$$\begin{aligned}\|\mu\|_{W^{-1/2,2}(\partial D)}&=\langle\mu,\mu\rangle^{1/2}_{W^{-1/2,2}(\partial D)}\\&=\sup_{v\in W^{1,2}(D)\setminus\{0\}}\frac{\langle w,v\rangle_{W^{1,2}(D)}}{\|v\|_{W^{1,2}(D)}}=\|w\|_{W^{1,2}(D)}.\end{aligned}$$

Well-posedness of (38)–(39) can be shown with a penalty version of Brezzi's splitting theorem, cf. e.g. [9]. In particular (u_β,λ_β) is bounded in $W^{1,2}(D)\times W^{-1/2,2}(\partial D)$ independent of β.
Since u_β depends Lipschitz-continuously on β, the error between the solution to (36)–(37) and (38)–(39) is $O(\beta)$; for a proof let us refer to [9, Ch. III, Thm 4.11, Cor. 4.15].

Lemma 6.11. *Let* $(u,\lambda),(u_\beta,\lambda_\beta) \in W^{1,2}(D)\times W^{-1/2,2}(\partial D)$ *be solutions to* (36)–(37) *and* (38)–(39)*, respectively. Then there exists a constant* C *independent of* β *such that*

$$\|u-u_\beta\|_{W^{1,2}(D)} + \|\lambda-\lambda_\beta\|_{W^{-1/2,2}(D)} \le C\beta.$$

Using $\mu = v/\beta$ with $v \in W^{1,2}(D)$ in (39) and adding the resulting equation to (38) yields the following reduced problem: Find $u_\beta \in W^{1,2}(D)$ such that

$$a(u_\beta, v) + \frac{1}{\beta}\int_{\partial D} u_\beta v \,\mathrm{d}\sigma = \ell(v) + \frac{1}{\beta}\int_{\partial D} gv \,\mathrm{d}\sigma \qquad \text{for all } v \in W^{1,2}(D).$$

This is a weak form of a Robin-type problem with boundary condition $n\cdot A\nabla u_\beta + \frac{1}{\beta}u_\beta = \frac{1}{\beta}g$ on ∂D.This method of relaxation of the Dirichlet boundary condition is widely known as the penalty method [4]. Let u_β^ε denote the diffuse approximation to u_β as defined in Section 6.1, i.e. u_β^ε satisfies

$$\begin{aligned}\int_\Omega A\nabla u_\beta^\varepsilon\cdot\nabla v + cu_\beta^\varepsilon v \,\mathrm{d}\omega^\varepsilon + \frac{1}{\beta}\int_\Omega u_\beta^\varepsilon v|\nabla\omega^\varepsilon|\,\mathrm{d}x \\ = \int_\Omega fv\,\mathrm{d}\omega^\varepsilon + \frac{1}{\beta}\int_\Omega gv|\nabla\omega^\varepsilon|\,\mathrm{d}x \qquad (40)\end{aligned}$$

for all $v \in W^{1,2}(D_\varepsilon;\omega^\varepsilon)$. Combining the estimates in Lemma 6.11 and Theorem 6.5, we have

$$\begin{aligned}\|u-u_\beta^\varepsilon\|_{W^{1,2}(D)} &\le \|u-u_\beta\|_{W^{1,2}(D)} + 2\|u_\beta - u_\beta^\varepsilon\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)} \\ &\le C(\beta + \frac{1}{\beta}\varepsilon^{\frac{1}{2}-\frac{1}{p}}) \qquad (41)\end{aligned}$$

for $p \le 2^*_\alpha$, and $u \in W^{1,p}(D)$. Choosing $\beta = \varepsilon^\sigma$, $\sigma > 0$, yields

$$\|u-u_\beta^\varepsilon\|_{W^{1,2}(D)} \le C(\varepsilon^\sigma + \varepsilon^{\frac{1}{2}-\frac{1}{p}-\sigma}).$$

Balancing the exponents on the right-hand side, we obtain the optimal choice $\sigma = \frac{1}{4}-\frac{1}{2p}$. The corresponding estimates are then given by the next theorems:

Theorem 6.12. *Let (C1)–(C2) hold. Moreover, assume that* $u \in W^{1,p}(D)$ *with* $2 \le p \le 2^*_\alpha$ *is a solution to* (34) *and* $u_\beta^\varepsilon \in W^{1,2}(D_\varepsilon;\omega^\varepsilon)$ *is a solution to* (40)*. Then for* $\beta = \varepsilon^\sigma$ *and* $\sigma = \frac{1}{4}-\frac{1}{2p}$ *there exists a constant* $C > 0$ *independent of* ε *such that*

$$\|u-u_\beta^\varepsilon\|_{W^{1,2}(D)} \le C\varepsilon^{\frac{1}{4}-\frac{1}{2p}}.$$

Theorem 6.13. *Let ∂D be of class C^∞, and let $f, g \in C^\infty(\overline{\Omega})$ and let (C1)–(C2) hold. Moreover, let $A \in C^\infty(\overline{\Omega})^{n\times n}$, $c \in C^\infty(\overline{\Omega})$, and let $u_\beta^\varepsilon \in W^{1,2}(D_\varepsilon;\omega^\varepsilon)$ denote the solution to (40), and let $u \in W^{1,2}(D)$ denote the solution to (34). Then, for $\beta = \varepsilon^\sigma$ and $\sigma = \frac{3}{4}$ there exists a constant C independent of ε such that*

$$\|u - u_\beta^\varepsilon\|_{W^{1,2}(D)} \leq C\varepsilon^{\frac{3}{4}}.$$

Remark 6.14. *Due to the regularity of ∂D, we can extend $u - u_\beta$ to $\mathbb{R}^n$ such that $\|u - u_\beta\|_{W^{1,2}(\mathbb{R}^n)} \leq C\|u - u_\beta\|_{W^{1,2}(D)}$ [1]. In view of (41) and the following chain of inequalities*

$$\begin{aligned}\|u - u_\beta\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)} &\leq \|u - u_\beta\|_{W^{1,2}(D_\varepsilon)} \leq \|u - u_\beta\|_{W^{1,2}(\mathbb{R}^n)} \\ &\leq C\|u - u_\beta\|_{W^{1,2}(D)} \leq C\beta\end{aligned}$$

the $W^{1,2}(D)$-norm in Theorem 6.12 and Theorem 6.13 can be replaced by $W^{1,2}(D_\varepsilon;\omega^\varepsilon)$-norm. Note, however, that for $v \in W^{1,2}(D_\varepsilon;\omega^\varepsilon)$ we have $v_{|D} \in W^{1,2}(D)$, but for the extension $\tilde{v}$ of $v_{|D}$ from D to $\mathbb{R}^n$ in general $\tilde{v}_{|D_\varepsilon} \neq v$.

Remark 6.15. *In order to obtain an analogous statement of Theorem 6.8 for the Dirichlet case, we would need an analog of Lemma 6.11 for the $W^{1,p}$-norm. Hence, for illustration let us assume that $\|u - u_\beta\|_{W^{1,p}(D)} \leq C\beta$ for $1 \leq p \leq \infty$. Moreover, by regularity of ∂D, we can assume stability of the extension of u and u_β to Ω, i.e. $\|u - u_\beta\|_{W^{1,p}(\Omega)} \leq C\|u - u_\beta\|_{W^{1,p}(D)}$. Then, we arrive at the estimate*

$$\begin{aligned}\|u - u_\beta^\varepsilon\|_{\mathcal{X}_p^\varepsilon} &\leq \|u - u_\beta\|_{\mathcal{X}_p^\varepsilon} + \|u_\beta - u_\beta^\varepsilon\|_{\mathcal{X}_p^\varepsilon} \\ &\leq C\|u - u_\beta\|_{W^{1,p}(\Omega)} + \|u_\beta - u_\beta^\varepsilon\|_{\mathcal{X}_p^\varepsilon} \\ &\leq C(\beta + \varepsilon^{1+\frac{1}{p}}/\beta) \leq C\varepsilon^{\frac{1}{2}+\frac{1}{2p}}\end{aligned}$$

using $\beta = \varepsilon^{\frac{1}{2}+\frac{1}{2p}}$ and Theorem 6.8. Assuming furthermore that the norms of $W^{1,1}(D_\varepsilon;\omega^\varepsilon)$ and $\mathcal{X}_1^\varepsilon$ are equivalent (uniform with respect to ε), and using continuity of the embedding $W^{1,1}(D) \hookrightarrow L^{\frac{n}{n-1}}(D)$ we infer that

$$\|u - u_\beta^\varepsilon\|_{L^{\frac{n}{n-1}}(D)} \leq C\varepsilon.$$

In particular for $n = 1$, we obtain $\|u - u_\beta^\varepsilon\|_{L^p(D)} = O(\varepsilon)$ for any $1 \leq p \leq \infty$, and for $n = 2$ we obtain $\|u - u_\beta^\varepsilon\|_{L^2(D)} = O(\varepsilon)$. The reader should compare this to the results of [14] where for $n = 1$ a rate $O(\varepsilon^{1-\delta})$ for any $\delta > 0$ in the L^∞-norm is shown. Moreover, in [34] an L^2-rate $O(\varepsilon)$ for Poisson's equation in three dimensions has been obtained numerically. There, it is also suggested to choose $\beta = \varepsilon$, which complies with our analysis. Let us note however that the diffuse domain method in [34] is somewhat different from ours.

6.3 Neumann boundary conditions

Consider the following second order elliptic equation with Neumann-type boundary condition: Find u such that

$$-\mathrm{div}(A\nabla u) + cu = f \quad \text{in } D, \tag{42}$$

$$n \cdot A\nabla u = g \quad \text{on } \partial D. \tag{43}$$

In order to obtain (weak) solutions to (42)–(43), let us consider the following weak formulation: Find $u \in W^{1,2}_\diamond(D) = \{v \in W^{1,2}(D) : \int_D v dx = 0\}$ such that

$$a(u,v) = \ell(v) \quad \text{for all } v \in W^{1,2}_\diamond(D), \tag{44}$$

with bilinear and linear form

$$a(u,v) = \int_D A\nabla u \cdot \nabla v + cuv \,\mathrm{d}x, \qquad \ell(v) = \int_D fv \,\mathrm{d}x + \int_{\partial D} gv \,\mathrm{d}\sigma.$$

In view of the usual Poincaré inequality for $W^{1,2}(D)$, the weak form (44) is well-posed under the assumptions (C1)-(C2).

Lemma 6.16. *Let (C1)–(C2) hold. Moreover, let $f \in L^2(D)$ and $g \in W^{1,2}(D)$. Then there exists a unique $u \in W^{1,2}_\diamond(D)$ satisfying (44), and there exists $C > 0$ such that*

$$\|u\|_{W^{1,2}(D)} \le C(\|f\|_{L^2(D)} + \|g\|_{L^2(\partial D)}).$$

The diffuse approximation of (44) is then: Find $u^\varepsilon \in W^{1,2}_\diamond(D_\varepsilon;\omega^\varepsilon) = \{v \in W^{1,2}(D_\varepsilon;\omega^\varepsilon) : \int_{D_\varepsilon} v d\omega^\varepsilon = 0\}$ such that

$$a^\varepsilon(u^\varepsilon, v) = \ell^\varepsilon(v) \quad \text{for all } v \in W^{1,2}_\diamond(D_\varepsilon;\omega^\varepsilon), \tag{45}$$

where the corresponding bilinear and linear form are given by

$$a^\varepsilon(u^\varepsilon, v) = \int_\Omega A\nabla u^\varepsilon \cdot \nabla v + cu^\varepsilon v \,\mathrm{d}\omega^\varepsilon,$$

$$\ell^\varepsilon(v) = \int_\Omega fv \,\mathrm{d}\omega^\varepsilon + \int_\Omega gv|\nabla\omega^\varepsilon| \,\mathrm{d}x.$$

Lemma 6.17. *Let (C1)–(C2) hold. Moreover, let $f \in L^2(D_\varepsilon;\omega^\varepsilon)$ and $g \in W^{1,2}(D_\varepsilon;\omega^\varepsilon)$. Then there exists a unique $u^\varepsilon \in W^{1,2}_\diamond(D_\varepsilon;\omega^\varepsilon)$ satisfying (45), and there exists $C > 0$ independent of ε such that*

$$\|u^\varepsilon\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)} \le C(\|f\|_{L^2(D_\varepsilon;\omega^\varepsilon)} + \|g\|_{W^{1,2}(D_\varepsilon;\omega^\varepsilon)}).$$

Proof. Continuity of a^ε and ℓ^ε with respect to the $W^{1,2}(D_\varepsilon;\omega^\varepsilon)$-topology is obvious. Coercivity of a^ε on $W^{1,2}_\diamond(D_\varepsilon;\omega^\varepsilon)$ is a direct consequence of the positivity of A and the Poincaré inequality, see Corollary 4.10. An application of the Lax-Milgram lemma yields the assertion. □

Having established existence of solutions to the Neumann problems, convergence results can now be derived as in the Robin case above when setting $b = 0$. We leave this to the reader. Let us mention that the restriction from $W^{1,2}(D_\varepsilon;\omega^\varepsilon)$ to the space $W^{1,2}_\diamond(D_\varepsilon;\omega^\varepsilon)$ is only necessary if $\inf_{x\in\Omega} c(x) = 0$. In this case the algebraic condition $\int_\Omega u\,\mathrm{d}\omega^\varepsilon = 0$ has to be treated with care in a numerical implementation. We do not want to go into details here, but let us refer the reader to [7]. If otherwise $\inf_{x\in\Omega} c(x) > 0$, we could equally well pose (45) in the space $W^{1,2}(D_\varepsilon;\omega^\varepsilon)$ and the implementational details are similar to those of the Robin case. We thus will not dwell on the Neumann case in our further discussion.

7 Numerical Results

In the following we report the results of numerical tests related to the above investigations used conformal first order finite elements. Our particular interest here is not the efficient solution of realistic problems, but rather to test the sharpness of error estimates in different situations by computational experiments. In order to have an "exact" solution u we solve the original problem with sharp interface on a very fine mesh such that the numerical error is negligible. Moreover, in the computation of the diffuse domain solution we make sure that the largest mesh parameter, i.e. $h_{\max}$, is for all computations smaller than ε^2 such that the numerical accuracy does not pollute the experimental order of convergence. The error $e^\varepsilon = u - u^\varepsilon$ will always be measured in the relative norms

$$\frac{\|u - u^\varepsilon\|_{W^{k,p}(D)}}{\|u\|_{W^{k,p}(D)}},$$

where $W^{0,p} = L^p$. We provide several log-log plots of errors vs. ε, which shall be comparable to the theoretical orders represented by lines in those plots, see Figure 1, 3, 6 and 8. Since the constants in the estimates cannot be made explicit, we have to fix one value and hence decide to plot the theoretical rates in all log-log plots such that they coincide with the experimental rates for the largest value of ε, see Figure 1, 3, 6 and 8.

For most simulations (Case A-D below) we work with the domain $D = \{(x_1, x_2) : x_1^2 + x_2^2 < 0.5\}$, which obviously satisfies all regularity requirements. The mesh representation of this domain D consists of $3,336,340$ vertices. The mesh representation of the domain D_ε is simply a scaling of the mesh representation of D with $1 + \varepsilon$. Finally we present an example with the domain $D = (0,1) \times (0,1)$, i.e. the unit square (Case E), which indicates that the same rates still hold for piecewise smooth domains. In all test cases we use the function S from Example 3.1 (i).

7.1 Case A: Robin BC with smooth parameters

In our first simulations, we consider the boundary value problem (27)-(28), with the smooth parameters

$$\begin{aligned} A(x_1,x_2) = c(x_1,x_2) = 1, \quad f(x_1,x_2) &= 10\sin(\pi x_1) - 5x_2^2, \\ g(x_1,x_2) = 0, \quad \alpha(x_1,x_2) &= 1. \end{aligned}$$

From Theorem 6.8, Corollary 6.9 and Remark 6.10, we expect the error e^ε to converge with a rate $O(\varepsilon^2)$ in $W^{1,1}(D)$ and $L^2(D)$ and a rate of $O(\varepsilon^{3/2})$ in $W^{1,2}(D)$. Furthermore we expect rates of order $O(\varepsilon)$ in $W^{1,\infty}(D)$. We mention that except the rate in $W^{1,2}(D)$ these expectations rely on assumptions we cannot verify rigorously. From Table 1 and the log-log plot of Figure 1 we observe that the numerical results reproduce these rates very accurately, indicating the sharpness of our estimates and the validity of the assumptions. In Figure 2, the solutions u and $u^\varepsilon|_D$ for the Robin boundary problem are presented. From a visual perspective, these solution are almost identical.

ε	$\|e^\varepsilon\|_{L^2}$	$\|e^\varepsilon\|_{W^{1,2}}$	$\|e^\varepsilon\|_{W^{1,1}}$	$\|e^\varepsilon\|_{W^{1,\infty}}$
2^{-1}	0.65477	0.75516	0.93812	0.68035
2^{-2}	0.19912 (1.71)	0.33747 (1.16)	0.38123 (1.30)	0.44465 (0.61)
2^{-3}	0.04953 (2.01)	0.12668 (1.41)	0.11738 (1.70)	0.24280 (0.87)
2^{-4}	0.01176 (2.07)	0.04475 (1.50)	0.03217 (1.87)	0.12474 (0.96)
2^{-5}	0.00281 (2.06)	0.01563 (1.52)	0.00846 (1.93)	0.06284 (0.99)

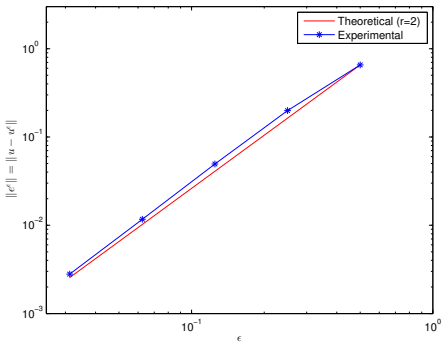
Table 1: The error $e^\varepsilon = u - u^\varepsilon$ for different norms in Case A. In paranthesis, we see the $\log_2$-ratio of $\frac{\|e_k^\varepsilon\|}{\|e_{k+1}^\varepsilon\|}$.

7.2 Case B: Robin BC with discontinuous A matrix

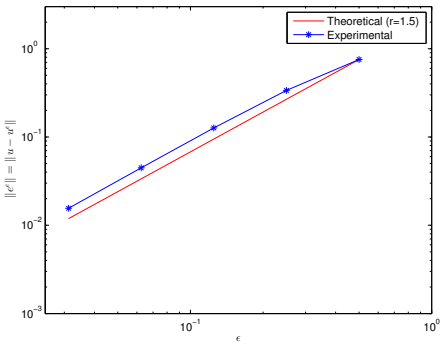
If the parameter A is no longer smooth, but instead $A \in L^\infty(\Omega)^{2\times 2}$, the assumptions for Theorem 6.8 are no longer satisfied. In the second example, we choose a discontinuous $A \in L^\infty(\Omega)^{2\times 2}$ as

$$A(x_1,x_2) = \begin{bmatrix} k_1(x_1,x_2) & 0 \\ 0 & k_2(x_1,x_2), \end{bmatrix}$$

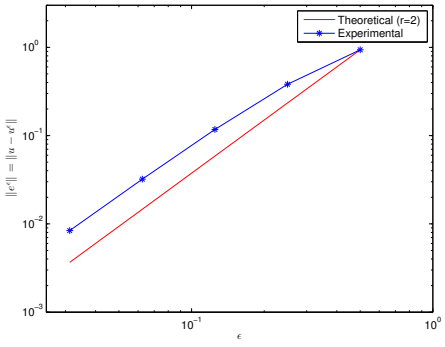
where k_1, k_2 are piecewise constant functions with a jump discontinuity close to ∂D. All other parameters are the same as in Case A.
From Table 2 and the log-log plot in Figure 3, we see that the convergence rate of the error is one order worse than in Case A. In particular, we obtain linear convergence in $W^{1,2}$, which is still better than the theoretical result of order $\varepsilon^{\frac{1}{2}}$ we obtain in the non-smooth case for $u \in W^{1,\infty}$. However we



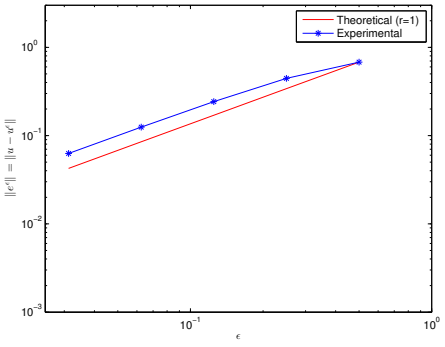
(a) Convergence in L^2-norm.



(b) Convergence in $W^{1,2}$-norm.

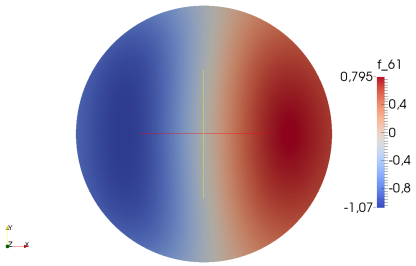


(c) Convergence in $W^{1,1}$-norm.

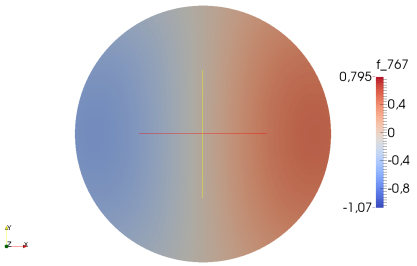


(d) Convergence in $W^{1,\infty}$-norm.

Figure 1: A log-log plot of the convergence rates in Case A. In each subplot we see the actual convergence rate (experimental), compared to the theoretical rate of order $O(\varepsilon^r)$. In subplots (a) and (c) $r = 2$, in (b) $r = 1.5$ and in (d) $r = 1$.



(a) The solution u of (30) for $\varepsilon = 2^{-1}$.



(b) The solution u^ε of (30) for $\varepsilon = 2^{-5}$.

Figure 2: Comparison of two diffuse domain solutions in Case A. The solution displayed in (b) is visually identical to the exact solution of (29).

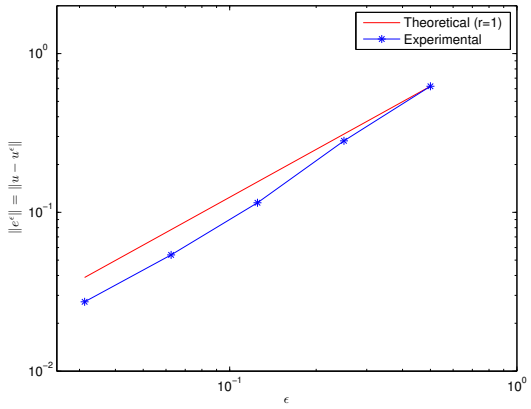


Figure 3: A log-log plot of the $W^{1,2}$-convergence in Case B. We see the actual convergence rate (experimental), compared to the theoretical rate of order $O(\varepsilon)$.

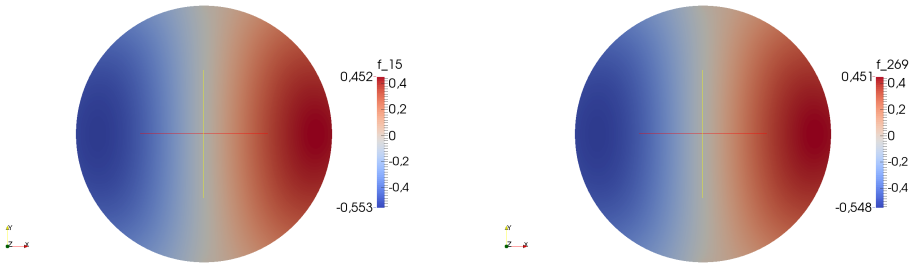
observe clearly the influence of non-smooth A on the convergence rate when comparing to case A. For the L^2-convergence, the rate is more inconsistent, as it appears to jump from quadratic to linear when $\varepsilon = 2^{-3}$. Although the convergence rate in $W^{1,2}$ is only linear in this case, a visual inspection of the solutions shown in Figure 4 reveals that the solution by the diffuse domain method is still almost identical to the exact solution for $\varepsilon = 2^{-5}$.

ε	$\|e^\varepsilon\|_{L^2}$	$\|e^\varepsilon\|_{W^{1,2}}$
2^{-1}	0.465577	0.622542
2^{-2}	0.127344 (1.87)	0.282197 (1.14)
2^{-3}	0.026850 (2.25)	0.114941 (1.30)
2^{-4}	0.014065 (0.93)	0.053956 (1.09)
2^{-5}	0.007958 (0.82)	0.027376 (0.98)

Table 2: The error $e^\varepsilon = u - u^\varepsilon$ for different norms in Case B. In paranthesis, we see the $\log_2$-ratio of $\frac{\|e^\varepsilon_k\|}{\|e^\varepsilon_{k+1}\|}$.

7.3 Case C: Robin BC with non-smooth parameters

Furthermore, in Case C, we also refine the mesh around the discontinuity to increase accuracy.
In Case A, we obtained a convergence rate in $W^{1,2}$ of order $O(\varepsilon^{3/2})$. Here, everything were smooth. In Case B, when working with a discontinuous matrix A, the convergence rate in $W^{1,2}$ drops to order $O(\varepsilon)$. If, however, the function f in (27) is in L^2, but not in L^p for $p \gg 2$, Theorem 6.5 yields $W^{1,2}$-convergence of, in worst case, order $O(\varepsilon^{1/3})$.



(a) The solution u of (29). (b) The solution u^ε of (30).

Figure 4: Comparison of exact solution and diffuse domain solution in Case B. Both solution restricted to D. Here $\varepsilon = 2^{-5}$.

To explore this issue numerically, we define

$$f(x_1, x_2) = \frac{1}{|x - y|^\mu}, \text{ where } y \in \partial D \text{ is fixed.} \tag{46}$$

Thus, we get that $f \in L^p(D)$ whenever $\frac{2}{\mu} > p$. All other parameters are given as in Case A. With a similar reasoning as in Lemma 6.3, using $n = 2$ and $\alpha = 1$, we expect a rate of convergence of $O(\varepsilon^{5/6-1/p})$.
In Figure 5, we see the convergence rate as a function of the parameter μ in (46). As expected, the convergence rate becomes worse when μ increases. The experimental rate deteriorates more, however, than the theory suggests. We believe this to be linked to the challenge of the numerical implementation of such a singular function. Although of practical importance, we find that dealing with this particular implementation issue is beyond the scope of this article, and leave it therefore to future research.

7.4 Case D: Dirichlet BC with smooth parameters

We will now study the diffuse domain method for a Dirichlet problem. More particularly, we will compare the solutions u and u^ε of (34) and (40), respectively. The parameters are given as in case A, with the exception

$$\alpha(x_1, x_2) = \frac{1}{\varepsilon^\sigma}$$

in order to realize the penalty method. From Theorem 6.13 and Remark 6.15, the choice of $\beta = \varepsilon^{-1}$ should provide a L^2-convergence of order $O(\varepsilon)$, whereas the choice of $\beta = \varepsilon^{-3/4}$ should yield a $W^{1,2}$-convergence of order $O(\varepsilon^{3/4})$.
In Table 3(a), we see the convergence rate of the error when $\beta = \varepsilon^{-3/4}$. As expected, the rate is of order $O(\varepsilon^{3/4})$ in $W^{1,2}$ norm. Furthermore, in Table 3(b), we obtain the expected linear convergence in L^2 norm. The rates can also be seen in Figure 6.

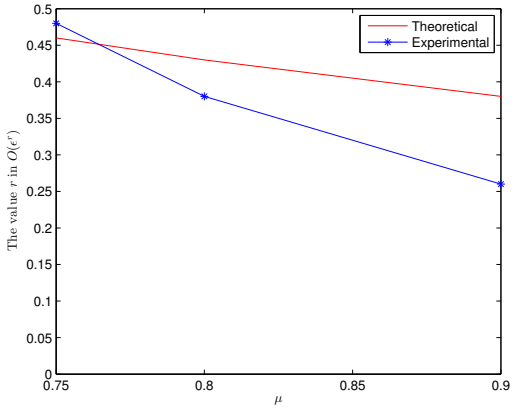


Figure 5: A plot of the $W^{1,2}$-convergence in Case C. The x-axis displays the parameter μ in (46), while the y-axis shows the theoretical and experimental convergence rate of corresponding function in L^p . The theoretical convergence rate follows from Theorem 6.5, and is of order $O(\varepsilon^{\frac{5}{6}-\frac{1}{p}})$.

In Figure 7, we see a solution of (34) and (40), respectively. Although the shape of the solution is visually similar, there is a much larger quantitative difference compared to Case A and B.

ε	$\|e^\varepsilon\|_{L^2}$	$\|e^\varepsilon\|_{W^{1,2}}$
2^{-1}	4.4342	2.4391
2^{-2}	2.3011 (0.95)	1.3949 (0.81)
2^{-3}	1.3653 (0.75)	0.8237 (0.76)
2^{-4}	0.8787 (0.64)	0.5091 (0.69)
2^{-5}	0.5669 (0.63)	0.3181 (0.68)

(a) $\sigma = 0.75$.

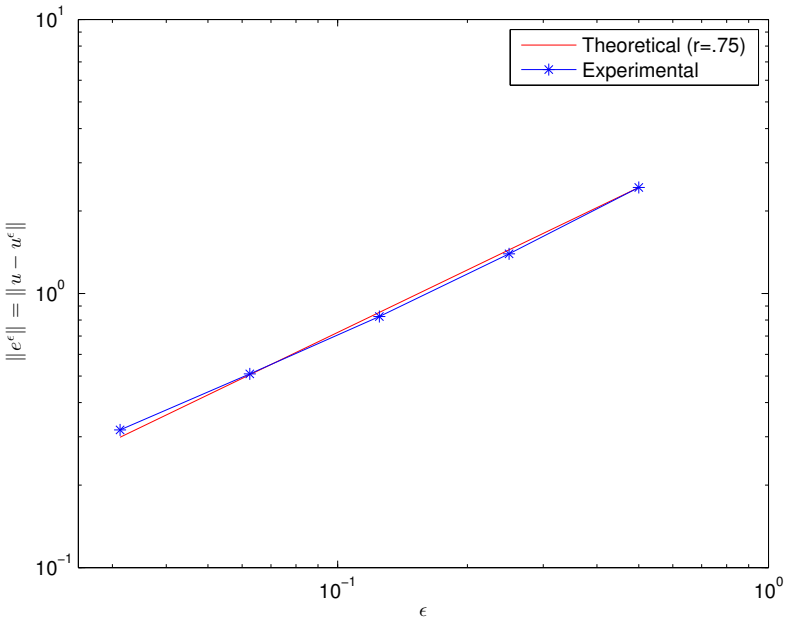
ε	$\|e^\varepsilon\|_{L^2}$	$\|e^\varepsilon\|_{W^{1,2}}$
2^{-1}	4.1093	2.2947
2^{-2}	1.8443 (1.16)	1.1888 (0.95)
2^{-3}	0.8689 (1.09)	0.6165 (0.95)
2^{-4}	0.4322 (1.01)	0.3318 (0.89)
2^{-5}	0.2179 (0.99)	0.1863 (0.83)

(b) $\sigma = 1.00$.

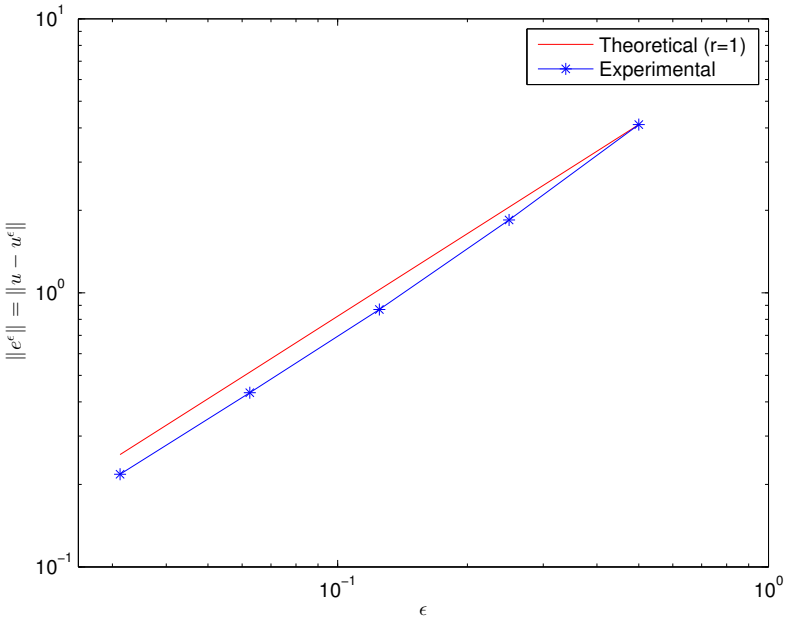
Table 3: The error $e^\varepsilon = u - u^\varepsilon$ for different norms. In paranthesis, we see the $\log_2$-ratio of $\frac{\|e^\varepsilon_k\|}{\|e^\varepsilon_{k+1}\|}$.

7.5 Case E: Dirichlet BC with smooth parameters

To guarantee a quadratic convergence in L^2, Theorem 6.8 requires the domain to be $C^{1,1}$. In this final example, we work with the mesh $D = (0,1) \times (0,1)$, which is only a Lipschitz domain. All parameters are otherwise identical to in Case A. We see from Figure 8 that the convergence rates are unchanged compared to case A despite the lower regularity of the domain. This gives some hope that our results can be extended to general Lipschitz domain or at least piecewise smooth domains, which remains an

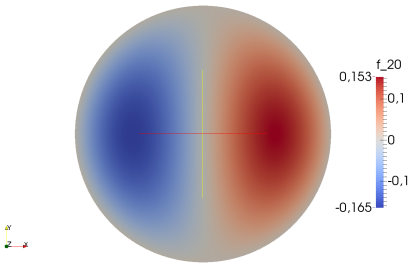


(a) $W^{1,2}$ convergence. $\sigma = 0.75$.

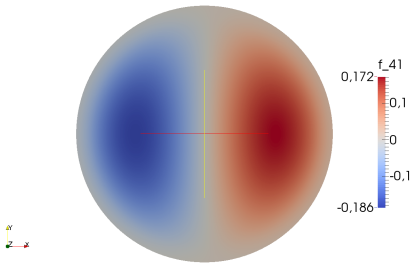


(b) L^2 convergence. $\sigma = 1.0$.

Figure 6: A log-log plot of the convergence rates in Case D. In each subplot we see the actual convergence rate (experimental), compared to the theoretical rate of order $O(\varepsilon^r)$.



(a) The solution u of (34).



(b) The solution u^ε of (40).

Figure 7: Comparison of exact solution and diffuse domain solution in Case D. Both solution restricted to D. Here $\varepsilon = 2^{-5}$ and $\sigma = 1$.

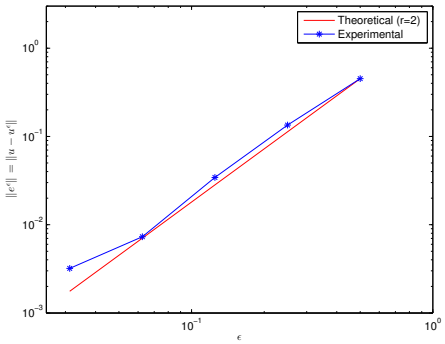
open question.

8 Conclusions

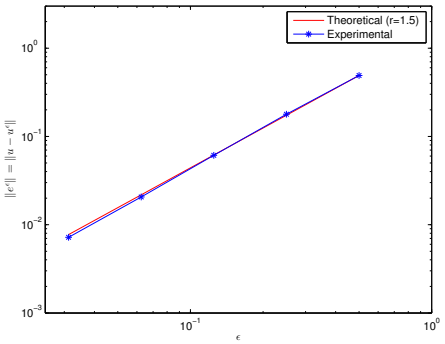
In this work we presented a systematic approach for deriving diffuse domain approaches for second order elliptic problems with usual type of boundary conditions. The advantage of our method is that based on standard variational formulations it readily leads to a relaxed variational formulation, which can be implemented easily, in a straight-forward manner. We presented a self-contained analysis of the error introduced by the diffuse domain method. Depending on the regularity of the data, we could rigorously prove convergence rates. These rates seem to be sharp as shown by numerical experiments. As a by-product of our analysis, we derived trace and embedding theorems as well as Poincaré inequalities for weighted Sobolev spaces which are stable with respect to the relaxation parameter ε. It remains open to fill a gap to transfer our quadratic convergence results to quadratic convergence results in the $L^2(D)$-norm which have been proposed in literature. Furthermore, a thorough analysis of numerical methods for the diffuse domain method is left for future work. We are optimistic that time-dependent problems could be treated in a similar manner, further modifications will be needed in the case of evolving surfaces.

Acknowledgements

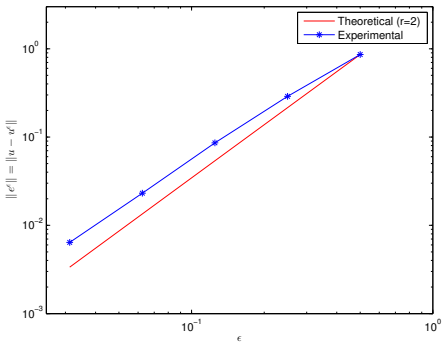
MB and MS acknowledge support by ERC via Grant EU FP 7 - ERC Consolidator Grant 615216 LifeInverse. MB acknowledges support by the German Science Foundation DFG via EXC 1003 Cells in Motion Cluster of Excellence, Münster, Germany. OLE acknowledges support by DAAD for his one year research stay at WWU Münster.



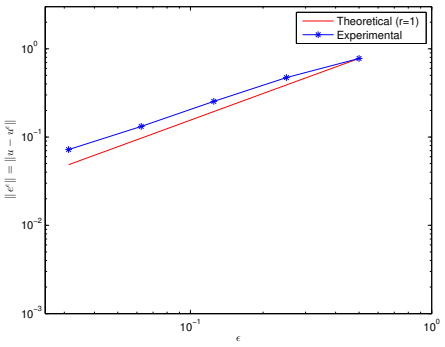
(a) Convergence in L^2-norm.



(b) Convergence in $W^{1,2}$-norm.



(c) Convergence in $W^{1,1}$-norm.



(d) Convergence in $W^{1,\infty}$-norm.

Figure 8: A log-log plot of the convergence rates in Case E. In each subplot we see the actual convergence rate (experimental), compared to the theoretical rate of order $O(\varepsilon^r)$.

References

[1] R. A. Adams. *Sobolev spaces*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975. Pure and Applied Mathematics, Vol. 65.

[2] S. Aland, J. Lowengrub, and A. Voigt. Two-phase flow in complex geometries: a diffuse domain approach. *CMES Comput. Model. Eng. Sci.*, 57(1):77–107, 2010.

[3] J. M. Arrieta, A. Rodríguez-Bernal, and J. D. Rossi. The best Sobolev trace constant as limit of the usual Sobolev constant for small strips near the boundary. *Proc. Roy. Soc. Edinburgh Sect. A*, 138(2):223–237, 2008.

[4] I. Babuška. The finite element method with penalty. *Math. Comp.*, 27:221–228, 1973.

[5] J. W. Barrett and C. M. Elliott. Fitted and unfitted finite-element methods for elliptic equations with smooth interfaces. *IMA J. Numer. Anal.*, 7(3):283–300, 1987.

[6] P. Bastian and C. Engwer. An unfitted finite element method using discontinuous Galerkin. *Internat. J. Numer. Methods Engrg.*, 79(12):1557–1576, 2009.

[7] P. Bochev and R. B. Lehoucq. On the finite element solution of the pure Neumann problem. *SIAM Rev.*, 47(1):50–66, 2005.

[8] A. Boulkhemair and A. Chakib. On the uniform Poincaré inequality. *Comm. Partial Differential Equations*, 32(7-9):1439–1447, 2007.

[9] D. Braess. *Finite elements*. Cambridge University Press, Cambridge, third edition, 2007. Theory, fast solvers, and applications in elasticity theory, Translated from the German by Larry L. Schumaker.

[10] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge*, 8(R-2):129–151, 1974.

[11] M. C. Delfour and J.-P. Zolésio. *Shapes and geometries*, volume 22 of *Advances in Design and Control*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 2011. Metrics, analysis, differential calculus, and optimization.

[12] H. Egger and M. Schlottbom. Analysis and regularization of problems in diffuse optical tomography. *SIAM J. Math. Anal.*, 42(5):1934–1948, 2010.

[13] S. Esedoḡlu, A. Rätz, and M. Röger. Colliding interfaces in old and new diffuse-interface approximations of Willmore-flow. *Commun. Math. Sci.*, 12(1):125–147, 2014.

[14] S. Franz, R. Gärtner, H.-G. Roos, and A. Voigt. A note on the convergence analysis of a diffuse-domain approach. *Comput. Methods Appl. Math.*, 12(2):153–167, 2012.

[15] R. Glowinski, T.-W. Pan, and J. Périaux. A fictitious domain method for Dirichlet problem and applications. *Comput. Methods Appl. Mech. Engrg.*, 111(3-4):283–303, 1994.

[16] J. B. Greer. An improvement of a recent Eulerian method for solving PDEs on general geometries. *J. Sci. Comput.*, 29(3):321–352, 2006.

[17] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Pitman, Boston, 1985.

[18] K. Gröger. A $W^{1,p}$-estimate for solutions to mixed boundary value problems for second order elliptic differential equations. *Math. Ann.*, 283(4):679–687, 1989.

[19] W. Hackbusch and S. A. Sauter. Composite finite elements for the approximation of PDEs on domains with complicated micro-structures. *Numer. Math.*, 75(4):447–472, 1997.

[20] T. Horiuchi. The imbedding theorems for weighted Sobolev spaces. *J. Math. Kyoto Univ.*, 29(3):365–403, 1989.

[21] A. Kufner. *Weighted Sobolev spaces*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1985. Translated from the Czech.

[22] K. Y. Lervag and J. Lowengrub. Analysis of the diffuse-domain method for solving PDEs in complex geometries. *arxiv:1407.7480v1*, 2014.

[23] R. J. LeVeque and Z. L. Li. The immersed interface method for elliptic equations with discontinuous coefficients and singular sources. *SIAM J. Numer. Anal.*, 31(4):1019–1044, 1994.

[24] X. Li, J. Lowengrub, A. Rätz, and A. Voigt. Solving PDEs in complex geometries: a diffuse domain approach. *Commun. Math. Sci.*, 7(1):81–107, 2009.

[25] F. Liehr, T. Preusser, M. Rumpf, S. Sauter, and L. O. Schwen. Composite finite elements for 3D image based computing. *Comput. Vis. Sci.*, 12(4):171–188, 2009.

[26] N. G. Meyers. An L^p-estimate for the gradient of solutions of second order elliptic divergence equations. *Annali della Scuola Normale Superiore di Pisa - Classe di Scienze*, 17(3):189–206, 1963.

[27] J. Nečas. *Direct methods in the theory of elliptic equations*. Springer Monographs in Mathematics. Springer, Heidelberg, 2012. Translated from the 1967 French original by Gerard Tronel and Alois Kufner, Editorial coordination and preface by Šárka Nečasová and a contribution by Christian G. Simader.

[28] B. Opic and A. Kufner. *Hardy-type inequalities*, volume 219 of *Pitman Research Notes in Mathematics Series*. Longman Scientific & Technical, Harlow, 1990.

[29] F. Otto, P. Penzler, A. Rätz, T. Rump, and A. Voigt. A diffuse-interface approximation for step flow in epitaxial growth. *Nonlinearity*, 17(2):477–491, 2004.

[30] J. Parvizian, A. Düster, and E. Rank. Finite cell method: h- and p-extension for embedded domain problems in solid mechanics. *Comput. Mech.*, 41(1):121–133, 2007.

[31] C. S. Peskin. Numerical analysis of blood flow in the heart. *J. Computational Phys.*, 25(3):220–252, 1977.

[32] A. Rätz. A new diffuse-interface model for step flow in epitaxial growth. *IMA Journal of Applied Mathematics*, 2014.

[33] A. Rätz, A. Voigt, et al. Pde's on surfaces—a diffuse interface approach. *Communications in Mathematical Sciences*, 4(3):575–590, 2006.

[34] M. G. Reuter, J. C. Hill, and R. J. Harrison. Solving PDEs in irregular geometries with multiresolution methods I: Embedded Dirichlet boundary conditions. *Comput. Phys. Commun.*, 183(1):1–7, 2012.

[35] K. E. Teigen, P. Song, J. Lowengrub, and A. Voigt. A diffuse-interface method for two-phase flows with soluble surfactants. *J. Comput. Phys.*, 230(2):375–393, 2011.

[36] H. Triebel. *Theory of function spaces. III*, volume 100 of *Monographs in Mathematics*. Birkhäuser Verlag, Basel, 2006.

6

Paper V - Diffuse Interface Methods for Inverse Problems: Case Study for an Elliptic Cauchy Problem

This paper is submitted for publication.

Diffuse Interface Methods for Inverse Problems: Case Study for an Elliptic Cauchy Problem

Martin Burger*, Ole Løseth Elvetun† and Matthias Schlottbom‡

June 18, 2015

Abstract

Many inverse problems have to deal with complex, evolving and often not exactly known geometries, e.g. as domains of forward problems modeled by partial differential equations. This makes it desirable to use methods which are robust with respect to perturbed or not well resolved domains, and which allow for efficient discretizations not resolving any fine detail of those geometries. For forward problems in partial differential equations methods based on diffuse interface representations gained strong attention in the last years, but so far they have not been considered systematically for inverse problems. In this work we introduce a diffuse domain method as a tool for the solution of variational inverse problems. As a particular example we study ECG inversion in further detail. ECG inversion is a linear inverse source problem with boundary measurements governed by an anisotropic diffusion equation, which naturally cries for solutions under changing geometries, namely the beating heart.

We formulate a regularization strategy using Tikhonov regularization and, using standard source conditions, we prove convergence rates. A special property of our approach is that not only operator perturbations are introduced by the diffuse domain method, but more important we have to deal with topologies which depend on a parameter ε in the diffuse domain method, i.e. we have to deal with ε-dependent forward operators and ε-dependent norms. In particular the appropriate function spaces for the unknown and the data depend on ε. This prevents to apply some standard convergence techniques for inverse problems, in particular interpreting the perturbations as data errors in the original problem does not yield suitable results. We consequently develop a novel approach based on saddle-point problems.

*Institute for Computational and Applied Mathematics, University of Münster, Einsteinstr. 62, 48149 Münster, Germany; Cells in Motion Cluster of Excellence, University of Münster. E-mail: burger@uni-muenster.de

†Dept. of Mathematical Sciences and Technology, Norwegian University of Life Sciences. E-mail: ole.elvetun@nmbu.no

‡Institute for Computational and Applied Mathematics, University of Münster, Einsteinstr. 62, 48149 Münster, Germany. E-mail: schlottbom@uni-muenster.de

The numerical solution of the problem is discussed as well and results for several computational experiments are reported. In particular investigations of convergence rates support our theoretical findings.

Keywords: Diffuse domain method, inverse problems, variational regularization, convergence analysis, ECG inversion, Cauchy problem.

AMS subject classifications: 35R30, 35J20, 65N85, 65K10

1 Introduction

Mathematical models based on differential and integral equations to be solved on complex or time-varying domains play an important role in many applications, in particular in biomedicine due to the complexity and inherent motion of living systems. A straight-forward approach towards the numerical solution of such problems is to resolve the geometries by building appropriate grids and subsequent computation on those e.g. via finite element or finite volume methods. Due to the high complexity of building grids and interpolation issues between different time steps several approaches have emerged that avoid the explicit resolution of the geometry and rather work on a fixed grid, either directly by adapting the discretization scheme (cf. [3, 15, 20]) or by implicitly representing the geometry in terms of characteristic functions, level set functions or diffuse interfaces (cf. [4, 6, 14, 18, 19, 17, 25]). In the latter approach the interface is encoded via a function φ^ε that takes values close to $+1$ in the interior and -1 in the exterior of the domain to be represented, with an interfacial layer of smooth transition, which has a size of order ε. This approach is highly motivated by Cahn-Hilliard and phase-field models in materials science (cf. [2, 9, 8]). Analogous issues related to complex geometry frequently and increasingly arise in many inverse problems, e.g. in medical imaging shapes are obtained from segmentation of an anatomical imaging via MR or CT and subsequently used for other inversion tasks such as emission tomography or electromagnetic inversion (like EEG, MEG, ECG, MCG). Diffuse interface methods have however hardly been considered (cf. [10]), and in particular their convergence analysis has not been worked out in relation to regularization methods, which introduce another small parameter. To be more precise consider canonical inverse problems of the form

$$A(u) = f, \tag{1}$$

where $A : \mathcal{X} \to \mathcal{Y}$ is the forward operator between function spaces and f are noisy data. Those are to be solved by variational regularization techniques, which consist in minimizing

$$J(u) = \|A(u) - f\|_{\mathcal{Y}}^q + \alpha\|u - u_*\|_{\mathcal{X}}^r, \tag{2}$$

with $q, r \geq 1$ and u_* being a prior for the variable u, potentially equal to zero. There are three potential dependencies on the domain D. The first as direct dependence of the operator upon D, e.g. via partial differential equations to be solved on D in order to evaluate A. The diffuse interface method will introduce an approximation of the form

$$J^\varepsilon(u) = \|A^\varepsilon(u) - f^\varepsilon\|_{\mathcal{Y}^\varepsilon}^q + \alpha\|u - u_*\|_{\mathcal{X}^\varepsilon}^r, \tag{3}$$

with appropriate perturbations of operator, data, and norms. In particular the last fact creates novel theoretical questions, since the topologies of the ε-dependent space might not be equivalent to the ones of the original spaces $\mathcal{X}$ and $\mathcal{Y}$ as we shall see below. The convergence analysis thus needs to go beyond the current state of the theory and in this paper we use a novel approach based on saddle-point formulations. We also mention that our analysis does not mainly target the case of $\varepsilon \to 0$ for fixed α, which could be derived with similar techniques as used here and in [7].

We mention that from a practical point of view there are further reasons that can make diffuse interface methods attractive. A quite peculiar property is that due to the ill-posedness of most inverse problems and the consequently limited resolution of regularization methods high frequency information is lost. Intuitively this should also concern fine details in the geometry, hence smearing out the geometry information might not harm the quality of reconstructions or even further stabilize the problem. Another aspect is uncertainty in geometries, which may concern the domain (e.g. from incorrect segmentations) as well as the measurement locations (e.g. electrode positions on the body surface in EEG and ECG). A diffuse interface that averages the model over different possible domain shapes seems hence more appropriate than an exact treatment of the interface. A detailed study of these aspects is left to future research.

In the construction of diffuse interface methods we follow the approach in [7]. During the whole paper we shall assume to have a representation of an unknown shape $D \subset \Omega$ via its signed distance function d_D, i.e.,

$$d_D(x) = \begin{cases} +\operatorname{dist}(x, \partial D) & \text{if } x \in \Omega \setminus D, \\ -\operatorname{dist}(x, \partial D) & \text{if } x \in D. \end{cases} \tag{4}$$

The diffuse interface is then constructed via

$$\varphi^\varepsilon = S\left(-\frac{d_D}{\varepsilon}\right) \tag{5}$$

for $\varepsilon > 0$ small and S being a sigmoidal function, i.e., increasing with $\lim_{t\to\pm\infty} S(t) = \pm 1$. As ε tends to zero, S converges to the sign function, and hence φ^ε formally converges to

$$\varphi^0(x) = \begin{cases} -1 & \text{if } x \in \Omega \setminus D, \\ +1 & \text{if } x \in D. \end{cases} \tag{6}$$

Indeed this convergence can easily be made rigorous in L^p-spaces. In this work we use the sigmoidal function $S : \mathbb{R} \to \mathbb{R}$ defined by $S(t) = t/|t|$ for $|t| \geq 1$ and $S(t) = t$ for $|t| < 1$; more general choices are allowed and we refer the reader to [7]. Note that the support of $\nabla\varphi^\varepsilon$ is restricted to an ε-neighborhood of ∂D and that φ^ε is a Lipschitz-continuous function bounded by ± 1.
In order to obtain a representation with diffuse interfaces, we mainly need to discuss the approximation of integrals over the domain and its boundary. With such we can obviously treat most relevant issues: integral equations, inverse problems for partial differential equations via weak formulations, data fidelities and regularization terms in variational regularization methods. The only relevant case that needs additional considerations seems to be the different use of tangential and normal derivatives on curves or surfaces, which we postpone to future considerations. The key idea to approximate such integrals is a weighted averaging of the integrals on $\{d_D < t\}$ instead of the original domain $\{d_D < 0\}$ only (and similar for boundary integrals). Since $\frac{1}{2\varepsilon}S'(\frac{\cdot}{\varepsilon})$ approximates a concentrated distribution at zero, we expect

$$\begin{aligned}\int_D g(x)\,\mathrm{d}x &= \int_{\{d_D<0\}} g(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{1}{2\varepsilon}S'(-\frac{t}{\varepsilon}) \int_{\{d_D<0\}} g(x)\,\mathrm{d}x\,\mathrm{d}t \\ &\approx \int_{-\infty}^{\infty} \frac{1}{2\varepsilon}S'(-\frac{t}{\varepsilon}) \int_{\{d_D<t\}} g(x)\,\mathrm{d}x\,\mathrm{d}t \\ &= \frac{1}{2}\int_{-1}^{1} \int_{\{\varphi^\varepsilon>s\}} g(x)\,\mathrm{d}x\,\mathrm{d}s,\end{aligned}$$

where we have used the substitution $s = S(-\frac{t}{\varepsilon})$ in the last term. Now the layer cake-representation can further be used for given integrable g to rewrite

$$\int_{-1}^{1} \int_{\{\varphi^\varepsilon>s\}} g(x)\,\mathrm{d}x\,\mathrm{d}t = \int_\Omega \int_{-1}^{\varphi^\varepsilon(x)} \mathrm{d}s g(x)\,\mathrm{d}x = \int_\Omega (1+\varphi^\varepsilon)(x) g(x)\,\mathrm{d}x.$$

By an analogous computation we obtain for the boundary integral

$$\int_{\partial D} g(x)\,\mathrm{d}\sigma(x) \approx \frac{1}{2}\int_{-1}^{1} \int_{\{\varphi^\varepsilon=s\}} g(x)\,\mathrm{d}\sigma(x)\,\mathrm{d}s,$$

which can be simplified via the co-area formula to

$$\int_{-1}^{1} \int_{\partial\{\varphi^\varepsilon=s\}} g(x)\,\mathrm{d}\sigma(x)\,\mathrm{d}t = \int_\Omega g(x)|\nabla\varphi^\varepsilon(x)|\,\mathrm{d}x.$$

Detailed convergence results for these kind of integrals can be found in [7] and are recalled in the appendix.

Thus, integral functionals in (2) of the form

$$\mathcal{F}_{dom}(v) = \int_D \Psi(v, \nabla v, \dots, \nabla^m v)\,\mathrm{d}x \tag{7}$$

are approximated in a straight-forward way as

$$\mathcal{F}^{\varepsilon}_{dom}(v) = \int_\Omega \Psi(v, \nabla v, \dots, \nabla^m v)(1+\varphi^{\varepsilon})\,\mathrm{d}x. \tag{8}$$

Functionals on surfaces are less straight-forward with the exception of simple L^p-type regularization functional

$$\mathcal{F}_{bound}(v) = \int_{\partial D} \Psi(x, v)\,\mathrm{d}\sigma(x),$$

which have an obvious approximation

$$\mathcal{F}^{\varepsilon}_{bound}(v) = \int_\Omega \Psi(\cdot, v)|\nabla\varphi^{\varepsilon}(x)|\,\mathrm{d}x.$$

Gradient or higher-order derivative based regularization on surfaces is usually formulated in terms of tangential derivatives, whose diffuse approximation solely based on φ^{ε} is more involved. In this paper we will however restrict our attention to L^2-norms on the boundary of a domain, which can be approximated as $\mathcal{F}_{bound}$ above. From the construction we see however that the diffuse version of an L^2-norm (defined as the square root of $\mathcal{F}_{bound}$ with square Ψ) has an important topological difference to the L^2-norm on the sharp interface. Note that the latter roughly corresponds to an $H^{1/2}$-norm on the domain via trace theorems, hence the diffuse norm induces a weaker topology.

In the remainder of the paper we work out the convergence analysis of the diffuse interface approximation (3) in the example of ECG inversion, i.e. the solution of an elliptic Cauchy problem. This problem is well-studied on the one-hand from a theoretical point of view, but on the other hand leaves a clear practical challenge of efficient solution on different complex domains (moving hearts). More importantly, it includes a lot of the potential challenges for the convergence analysis: Both the unknown as well as the data are functions on parts of the boundary to be approximated by diffuse interfaces and the forward operator is also defined via a partial differential equation on the (diffuse) domain. We discuss the problem and its diffuse approximation in Section 2, before we proceed to the convergence analysis in Section 3. We show that the diffuse regularized solution converges to the correct solution as α, ε and the noise level δ tend to zero under standard conditions on α and roughly for $\varepsilon \sim \alpha$ (or some higher power of α). In the case of correct solutions satisfying a standard source condition (cf. [12]) and a standard choice $\alpha \sim \delta$ we obtain an optimal convergence rate if $\varepsilon \sim \delta^{2/3}$.

This confirms our intuition that ε can be chosen rather large for inverse problems in presence of noise. Finally we discuss the numerical solution of the problem in Section 5 and provide a collection of experiments, whose results support our theory respectively indicate that one might obtain even better convergence rates with respect to ε.

2 Motivating Example: ECG Inversion

In order to clarify the application of the diffuse domain method to the solution of an inverse problem, we study the following setup encountered in the reconstruction of epicardial potentials from ECG body surface potential measurements. Given data f, which are samples of the potential v (more precisely its Dirichlet trace on the body surface ∂B) we want to reconstruct the epicardial potential, i.e., the trace of v on ∂H, where $H \subset B$ is the heart volume. Here we use a so-called flux-based formulation, i.e., we use the Neumann boundary value u on ∂H as the unknown for the inversion, i.e., the forward model in weak form is

$$\int_D M\nabla v \cdot \nabla w \,\mathrm{d}x = \int_{\partial H} uw \,\mathrm{d}\sigma \quad \text{for all } w \in H^1_\diamond(D). \tag{9}$$

with $D = B \setminus \overline{H}$ and

$$H^1_\diamond(D) = \{w \in H^1(D) : \int_{\partial H} w \,\mathrm{d}\sigma = 0\}.$$

This formulation has been found to be quite appealing in the ECG-inversion problem, in particular when variational regularization is formulated on u rather than the Dirichlet trace of v (cf. [13, 16, 26]). The epicardial potential can be computed subsequently from the forward model. Note that (9) is the weak formulation of the anisotropic Laplace equation $\nabla \cdot (M\nabla v) = 0$ with Neumann boundary conditions, with zero flux on ∂B. The latter is natural due to the insulation of the body.
In the whole manuscript we will assume the following ellipticity condition: There exists a constant $m > 0$ such that

$$m|\xi|^2 \leq \xi \cdot M(x)\xi \leq \frac{1}{m}|\xi|^2 \quad \text{for all } x, \xi \in \mathbb{R}^n. \tag{10}$$

Moreover, we will always assume the following regularities: $\partial D \in C^{3,1}$, $M \in W^{2,\infty}(\Omega)$ and $v \in W^{3,\infty}(D)$ being the solution of (9). Thus, $n \cdot M\nabla v \in W^{2,\infty}(\partial D)$. These regularity assumptions can be weakened at the cost of worse approximation properties of the diffuse domain method, see some remarks below and [7].

Lemma 2.1. *Let* (10) *hold. Then, for any* $u \in L^2(\partial H)$*, there exists a unique* $v \in H^1_\diamond(D)$ *such that* (9) *holds. In particular, there exists a constant* $C > 0$ *such that*

$$\|v\|_{H^1(D)} \leq C\|u\|_{L^2(\partial H)}.$$

Proof. Due to the Poincaré inequality the bilinear form on the left-hand side of (9) defines an inner product on $H^1_\diamond(D)$. For $u \in L^2(\partial H)$ the right-hand side of (9) defines a bounded linear functional on $H^1_\diamond(D)$. An application of the Lax-Milgram lemma yields the assertion. □

2.1 Forward map and inverse problem.

We define a linear operator

$$F : L^2(\partial H) \to L^2(\partial B), \quad Fu = v_{|\partial B} \tag{11}$$

with $v \in H^1_\diamond(D)$ being the solution to (9) with $u \in L^2(\partial H)$. The inverse problem we are concerned with is the following. For given $f \in L^2(\partial B)$ determine $u \in L^2(\partial H)$ such that

$$Fu = f \qquad \text{in } L^2(\partial B). \tag{12}$$

The following lemma collects some basic properties of the forward map F.

Lemma 2.2. *The forward map* $F : L^2(\partial H) \to L^2(\partial B)$ *defined by* (11) *is linear, injective, bounded and compact.*

Proof. Linearity is obvious. Compactness, and hence boundedness, follows from compactness of the trace operator $H^1(D) \to L^2(\partial B)$ and Lemma 2.1. To show injectivity, let $u_1, u_2 \in L^2(\partial H)$ such that $Fu_1 = f = Fu_2$, and denote by v_1, v_2 the corresponding solutions to (9). Then the difference $w = v_1 - v_2$ is a weak solution to the Cauchy problem

$$-\mathrm{div}(M\nabla w) = 0 \quad \text{in } D, \qquad n \cdot M\nabla w = 0 \text{ on } \partial B, \quad w = 0 \text{ on } \partial B.$$

Since M is Lipschitz, the Cauchy problem is uniquely solvable [23], i.e., $w = 0$ and $u_1 = u_2$. □

In view of Lemma 2.2 and since it is easy to see that the range of F is infinite-dimensional, the inverse problem (12) is ill-posed, and some sort of regularization is needed for a stable inversion of (12). In the whole manuscript, we denote by $f^\dagger, v^\dagger$ and $u^\dagger$ the exact data and solutions respectively.

2.2 Variational Regularization with Sharp Interfaces

As basic regularization method we consider the following Tikhonov type functional

$$J(u,v) = \frac{1}{2}\|v - f^\delta\|^2_{L^2(\partial B)} + \frac{\alpha}{2}\|u\|^2_{L^2(\partial H)} \quad \text{subject to (9)}, \tag{13}$$

where $f^\delta \in L^2(\partial B)$ represents noisy data for which we assume that

$$\|f^\dagger - f^\delta\|_{L^2(\partial B)} \leq \delta. \tag{14}$$

As pointed out in the introduction, in applications we have in mind the sharp interfaces ∂B and ∂H are not known exactly, and we aim at employing the diffuse integrals introduced above. The quadratic case seems to be sufficient to understand the main difficulties arising from the diffuse approximation, extensions to other L^p-norms can be made with analogous arguments as in the sharp interface case.

Remark 2.3. *Considering the reduced functional $\hat{J}(u) = J(u, F(u))$, which is quadratic and strictly convex, we obtain from [12, Thm 5.2] that the minimizers $u_{\alpha,\delta}$ of $\hat{J}$ with $f^\dagger$ replaced by f^δ converge to $u^\dagger$ in $L^2(\partial H)$ as long as $u^\dagger \in L^2(\partial H)$, $\|f^\dagger - f^\delta\|_{L^2(\partial B)} \leq \delta$ and $\alpha \to 0$ is chosen such that $\delta^2/\alpha \to 0$ as $\delta \to 0$, i.e., $\lim_{\delta\to 0} u_{\alpha,\delta} = u^\dagger$.*

2.3 Variational Regularization with Diffuse Interface

In the following we discuss a diffuse approximation of the variational problems introduced above. In order to distinguish the two different parts of the boundary $\partial D = \partial B \cup \partial H$ we choose a weight γ_H that equals one in a neighborhood of ∂H and zero in a neighborhood of ∂B. Vice versa, we choose a second weight γ_B, which equals one in a neighborhood of the measurement locations on ∂B and vanishes in a neighborhood of ∂H.

2.3.1 Sobolev Spaces

To define a suitable function space, let us introduce the scalar product

$$\langle v, w\rangle_{\mathcal{H}^\varepsilon} = \langle \nabla v, \nabla w\rangle_{\omega^\varepsilon} + \langle v, w\rangle_{\omega^\varepsilon} = \int_\Omega (\nabla v \cdot \nabla w + vw)\omega^\varepsilon \,\mathrm{d}x,$$

where $\omega^\varepsilon = (1 + \varphi^\varepsilon)/2$, and the corresponding weighted Sobolev space defined by

$$\mathcal{H}^\varepsilon := \{v \in L^2(\Omega) | \|v\|^2_{\mathcal{H}^\varepsilon} = \langle v, v\rangle_{\mathcal{H}^\varepsilon} < \infty\}.$$

Note that we tacitly identify functions v and w if $v = w$ on $\mathrm{supp}(\omega^\varepsilon)$ in order to make $\|\cdot\|_{\mathcal{H}^\varepsilon}$ a norm. Moreover, we denote by $L^p(\omega^\varepsilon) = L^p(\Omega; \omega^\varepsilon)$ and

$W^{k,p}(\omega^\varepsilon) = W^{k,p}(\Omega, \omega^\varepsilon)$ the corresponding weighted Lebesgue and Sobolev spaces; see [7] for details. In particular $\mathcal{H}^\varepsilon = W^{1,2}(\omega^\varepsilon)$. We will also write $L^p(\tilde{\Omega}; \gamma)$ with some weighting function γ and $\tilde{\Omega} \subset \Omega$ to denote the corresponding weighted Lebesgue space. One observes that due to the properties of ω^ε, we have $\sqrt{2}\|v\|_{\mathcal{H}^\varepsilon} \geq \|v\|_{H^1(D)}$. Thus, any uniform estimate and convergence in the norm of $\mathcal{H}^\varepsilon$ can be transfered immediately to the norm of v in $H^1(D)$, which is a relevant quantity to understand the approximation properties; for further details on the spaces $\mathcal{H}^\varepsilon$ see also [7]. For the interface variable u we consider the space $\mathcal{U}^\varepsilon = L^2(\gamma_H|\nabla\omega^\varepsilon|)$ with corresponding inner product $\langle\cdot,\cdot\rangle_{\mathcal{U}^\varepsilon}$; and for the measurements f we use $\mathcal{M}^\varepsilon = L^2(\gamma_B|\nabla\omega^\varepsilon|)$ with corresponding inner product $\langle\cdot,\cdot\rangle_{\mathcal{M}^\varepsilon}$; i.e. for $u, q \in \mathcal{U}^\varepsilon$ and $f, v \in \mathcal{M}^\varepsilon$

$$\langle u, q\rangle_{\mathcal{U}^\varepsilon} = \int_\Omega uq|\nabla\omega^\varepsilon|\gamma_H \,\mathrm{d}x, \qquad \langle v, f\rangle_{\mathcal{M}^\varepsilon} = \int_\Omega vf|\nabla\omega^\varepsilon|\gamma_B \,\mathrm{d}x.$$

As above, we identify functions $u, q \in \mathcal{U}^\varepsilon$ if $u = q$ on $\mathrm{supp}(|\nabla\omega^\varepsilon|\gamma_H)$. The diffuse trace lemma A.3 shows that the embedding $\mathcal{H}^\varepsilon \hookrightarrow \mathcal{U}^\varepsilon$ is continuous. For appropriate normalization, we will also consider space

$$\mathcal{H}^\varepsilon_\diamond = \{v \in \mathcal{H}^\varepsilon : \langle v, 1\rangle_{\mathcal{U}^\varepsilon} = 0\}. \tag{15}$$

As ∂D is smooth, there exists a continuous extension $E_{D,\Omega} : H^1(D) \to H^1(\Omega)$ [1], and we will write v instead of $E_{D,\Omega}v$ to evaluate functions in $H^1(D)$ in Ω.

2.3.2 Extensions constant off the interface

We consider extensions constant off the interfaces ∂H and ∂B, respectively. Note that for $0 < \varepsilon \leq \varepsilon_0$, with ε_0 sufficiently small, which we will assume throughout the paper, and for each $x \in \mathrm{supp}(|\nabla\omega^\varepsilon|)$ there exists a unique $\bar{x} \in \partial D$ such that $x = \bar{x} + d_D(x)n(\bar{x})$; see [11]. We can then define $E_H : L^2(\partial H) \to \mathcal{U}^\varepsilon$ by

$$E_H u(x) = \tilde{u}(x) = u(\bar{x}), \qquad x = \bar{x} + d_D(x)n(\bar{x}) \in \mathrm{supp}(\gamma_H|\nabla\omega^\varepsilon|),\ \bar{x} \in \partial H,$$

and similarly for the measurements, $E_B : L^2(\partial B) \to \mathcal{M}^\varepsilon$ given by

$$E_B f(x) = \tilde{f}(x) = f(\bar{x}), \qquad x = \bar{x} + d_D(x)n(\bar{x}) \in \mathrm{supp}(\gamma_B|\nabla\omega^\varepsilon|),\ \bar{x} \in \partial B.$$

If the context is clear, we will write in abuse of notation $\tilde{u}$ and $\tilde{f}$ instead of $E_H u$ and $E_B f$. Some properties of the extensions E_B and E_H are compiled in the appendix.

2.3.3 Diffuse forward operator

We approximate (9) via

$$\langle M\nabla v, \nabla w\rangle_{\omega^\varepsilon} = \langle u, w\rangle_{\mathcal{U}^\varepsilon} \quad \text{for all } w \in \mathcal{H}^\varepsilon_\diamond, \tag{16}$$

where $u \in \mathcal{U}^\varepsilon$. We have the following well-posedness result for (16); see [7, Lemma 6.17].

Lemma 2.4. *For each $u \in \mathcal{U}^\varepsilon$ there exist a unique $v \in \mathcal{H}^\varepsilon_\diamond$ verifying* (16) *and a constant $C > 0$ independent of ε such that*

$$\|v\|_{\mathcal{H}^\varepsilon} \leq C\|u\|_{\mathcal{U}^\varepsilon}.$$

In order to use $u^\dagger$ in (16), we will use the extension $\tilde{u}^\dagger = E_H u^\dagger \in \mathcal{U}^\varepsilon$. This will introduce errors quantified by the following

Lemma 2.5. *Let $v^\varepsilon \in \mathcal{H}^\varepsilon_\diamond$ be a solution to* (16) *with data u replaced by $E_H u^\dagger$. Then there exists $C > 0$ such that*

$$\|v^\dagger - v^\varepsilon\|_{\mathcal{H}^\varepsilon} \leq C\varepsilon^{3/2}\|v^\dagger\|_{W^{3,\infty}(D)}.$$

Proof. The difference $v^\varepsilon - v^\dagger$ satisfies

$$\langle M\nabla(v^\varepsilon - v^\dagger), \nabla w\rangle_{\omega^\varepsilon} = \langle \tilde{u}^\dagger, w\rangle_{\mathcal{U}^\varepsilon} - \langle M\nabla v^\dagger, \nabla w\rangle_{\omega^\varepsilon}.$$

Integration by parts and $-n|\nabla\omega^\varepsilon| = \nabla\omega^\varepsilon$ yields

$$\begin{aligned}\langle \tilde{u}^\dagger, w\rangle_{\mathcal{U}^\varepsilon} - \langle M\nabla v^\dagger, \nabla w\rangle_{\omega^\varepsilon} &= \langle \tilde{u}^\dagger - n\cdot M\nabla v^\dagger, w\rangle_{\mathcal{U}^\varepsilon} \\ &\quad - \langle n\cdot M\nabla v^\dagger, w\rangle_{\mathcal{M}^\varepsilon} - \langle \operatorname{div}(M\nabla v^\dagger), w\rangle_{\omega_\varepsilon}.\end{aligned}$$

To treat the first term we use $n\cdot M\nabla v^\dagger = u^\dagger$ on ∂H and Lemma A.4 (iv) to obtain

$$\langle E_H(n\cdot M\nabla v^\dagger) - n\cdot M\nabla v^\dagger, w\rangle_{\mathcal{U}^\varepsilon} \leq C\varepsilon^{3/2}\|v^\dagger\|_{W^{3,2}(\Omega)}\|w\|_{\mathcal{H}^\varepsilon}$$

for some $C > 0$. Since $n\cdot M\nabla v^\dagger = 0$ on ∂B, the second term can be treated similarly. To treat the third term we use $\operatorname{div}(M\nabla v^\dagger) = 0$ in D and Lemma A.4 (i) to obtain

$$|\langle \operatorname{div}(M\nabla v^\dagger), w\rangle_{\omega^\varepsilon}| \leq C\varepsilon^{3/2}\|\operatorname{div}(M\nabla v^\dagger)\|_{W^{1,\infty}(\Omega)}\|w\|_{\mathcal{H}^\varepsilon}.$$

The a priori estimate of Lemma 2.4 yields the assertion. □

Since in applications we have in mind both ∂B and ∂H are unknown or difficult to approximate, we will employ diffuse approximations of ∂B and ∂H. Hence, we are concerned with solving the following (diffuse) operator equation

$$F^\varepsilon u = \tilde{f}^\delta \quad \text{in } \mathcal{M}^\varepsilon, \tag{17}$$

where $F^\varepsilon : \mathcal{U}^\varepsilon \to \mathcal{M}^\varepsilon$ is a bounded linear operator mapping u onto the diffuse trace of the solution v of (16). The data $\tilde{f}^\delta = E_B f^\delta$ is obtained by extending the measured data f^δ. In view of the possible extensions of the

interface data u and f, there are of course many different possibilities to define a forward operator. Since these investigations will be similar to ours, we leave the modifications to the reader. Notice that, for each $\varepsilon > 0$ fixed, the injection $\mathcal{H}^\varepsilon \hookrightarrow \mathcal{M}^\varepsilon$ is compact, and hence (17) is ill-posed as well.
As E_B is bounded, see Lemma A.1, measuring in the weaker diffuse interface norm will not alter the noise level significantly, i.e.,

$$\|E_B f^\dagger - E_B f^\delta\|_{\mathcal{M}^\varepsilon} \leq C(\varepsilon)\delta, \tag{18}$$

with $C(\varepsilon) \to 1$ as $\varepsilon \to 0$. Using the diffuse domain method as underlying governing equation will however have an impact, which might be interpreted as an operator perturbation, namely

$$\|F^\varepsilon E_H u^\dagger - E_B f^\delta\|_{\mathcal{M}^\varepsilon} \leq C(\delta + \varepsilon^{3/2}).$$

The latter estimate is a direct consequence of the triangle inequality and Lemma 2.5. The Tikhonov functional (13) is approximated by the following functional

$$J^\varepsilon(u, v) = \frac{1}{2}\|v - \tilde{f}^\delta\|^2_{\mathcal{M}^\varepsilon} + \frac{\alpha}{2}\|u\|^2_{\mathcal{U}^\varepsilon} \quad \text{subject to (16).} \tag{19}$$

Note that we not only have to deal with perturbed forward operators but also with perturbed data misfit and regularization functionals. As the diffuse boundary norms are weaker than their counterparts for the sharp interfaces, this choice of topologies makes our investigations non-standard and requires adapted arguments to be detailed in the next section.

3 Analysis of the Diffuse Domain Regularization

In the following we provide an analysis of the variational models with diffuse interfaces. We begin with the existence of minimizers of (19) by investigating the associated saddle-point problem. Then we show stability and convergence of minimizers of the diffuse Tikhonov functional. Under a standard source condition we then also obtain convergence rates.

3.1 Saddle-Point Formulation

In the following we consider variations of the Lagrangian corresponding to (19)

$$L^\varepsilon(u, v, p) = J^\varepsilon(u, v) + \langle M\nabla v, \nabla p\rangle_{\omega^\varepsilon} - \langle u, p\rangle_{\mathcal{U}^\varepsilon}. \tag{20}$$

Therefore, let us define two bilinear forms, namely $a^\varepsilon : (\mathcal{U}^\varepsilon \times \mathcal{H}^\varepsilon_\diamond) \times (\mathcal{U}^\varepsilon \times \mathcal{H}^\varepsilon_\diamond) \to \mathbb{R}$ given by

$$a^\varepsilon(u, v; q, w) = \langle v, w\rangle_{\mathcal{M}^\varepsilon} + \alpha\langle u, q\rangle_{\mathcal{U}^\varepsilon},$$

and $b^\varepsilon : (\mathcal{U}^\varepsilon \times \mathcal{H}^\varepsilon_\diamond) \times \mathcal{H}^\varepsilon_\diamond \to \mathbb{R}$ given by

$$b^\varepsilon(u, v; p) = \langle M\nabla v, \nabla p\rangle_{\omega^\varepsilon} - \langle u, p\rangle_{\mathcal{U}^\varepsilon}.$$

Saddle-points of L^ε are then characterized as solutions of

$$\begin{array}{lcll} a^\varepsilon(u, v; q, w) + b^\varepsilon(q, w; p) & = & f^\varepsilon(q, w) & \text{for all } (q, w) \in \mathcal{U}^\varepsilon \times \mathcal{H}^\varepsilon_\diamond, \\ b^\varepsilon(u, v; r) & = & g^\varepsilon(r) & \text{for all } r \in \mathcal{H}^\varepsilon_\diamond. \end{array} \tag{21}$$

Here, we use the linear functionals $g^\varepsilon : \mathcal{H}^\varepsilon_\diamond \to \mathbb{R}$, $g^\varepsilon(r) = 0$, and $f^\varepsilon : \mathcal{U}^\varepsilon \times \mathcal{H}^\varepsilon_\diamond \to \mathbb{R}$, $f^\varepsilon(q, w) = \langle \tilde{f}^\delta, w\rangle_{\mathcal{M}^\varepsilon}$. For the analysis of the saddle-point problem, let us define

$$\|(u, v)\|^2_\alpha = \alpha(\|u\|^2_{\mathcal{U}^\varepsilon} + \|\nabla v\|^2_{L^2(\omega^\varepsilon)}) + \|v\|^2_{\mathcal{M}^\varepsilon},$$

which is a norm equivalent to the natural norm on $\mathcal{U}^\varepsilon \times \mathcal{H}^\varepsilon_\diamond$ for fixed $\alpha > 0$; cf. Lemma A.5.
Let us first collect some basic properties of the saddle-point problem and the associated bilinear forms:

Lemma 3.1 (Continuity)**.** *Let $0 < \alpha \leq \alpha_0$. Then there exists a constant C_c independent of ε and α such that*

$$|a^\varepsilon(u, v; q, w)| \leq C_c\|(u, v)\|_\alpha\|(q, w)\|_\alpha \quad \text{and}$$
$$|b^\varepsilon(u, v; p)| \leq \frac{1}{\sqrt{\alpha}} C_c\|(u, v)\|_\alpha\|p\|_{\mathcal{H}^\varepsilon}$$

for all $(u, v), (q, w) \in \mathcal{U}^\varepsilon \times \mathcal{H}^\varepsilon_\diamond$ and $p \in \mathcal{H}^\varepsilon_\diamond$.

Proof. The estimates follow from Lemma A.3 and a standard Cauchy-Schwarz argument. □

Lemma 3.2 (Kernel ellipticity)**.** *Let $0 < \alpha \leq \alpha_0$. Then there exists a constant C_e independent of ε and α such that*

$$a^\varepsilon(u, v; u, v) \geq C_e\|(u, v)\|^2_\alpha \tag{22}$$

for all $(u, v) \in \mathcal{U}^\varepsilon \times \mathcal{H}^\varepsilon_\diamond$ such that $b^\varepsilon(u, v; v) = 0$.

Proof. Using $b^\varepsilon(u, v; v) = 0$ we obtain for any $\kappa > 0$

$$\begin{aligned} a^\varepsilon(u, v; u, v) &= a^\varepsilon(u, v; u, v) + \kappa b^\varepsilon(u, v; v) \\ &\geq \|v\|^2_{\mathcal{M}^\varepsilon} + \alpha\|u\|^2_{\mathcal{U}^\varepsilon} + \kappa m\|\nabla v\|^2_{L^2(\omega^\varepsilon)} - \kappa\|u\|_{\mathcal{U}^\varepsilon}\|v\|_{\mathcal{U}^\varepsilon} \\ &\geq \|v\|^2_{\mathcal{M}^\varepsilon} + \frac{\alpha}{2}\|u\|^2_{\mathcal{U}^\varepsilon} + \kappa m\|\nabla v\|^2_{L^2(\omega^\varepsilon)} - \frac{\kappa^2}{2\alpha}\|v\|^2_{\mathcal{U}^\varepsilon}, \end{aligned}$$

where we have used (10) and Young's inequality. With Lemma A.3 and Lemma A.5 there exists a constant $c > 0$ independent of ε such that

$$\|v\|^2_{\mathcal{U}^\varepsilon} \leq c(\|\nabla v\|^2_{L^2(\omega^\varepsilon)} + \|v\|^2_{\mathcal{M}^\varepsilon}).$$

Increasing c if necessary, we may assume that $c \geq \alpha_0 m^2$. Hence, we arrive at the estimate

$$a^\varepsilon(u,v;u,v) \geq \|v\|^2_{\mathcal{M}^\varepsilon} + \frac{\alpha}{2}\|u\|^2_{\mathcal{U}^\varepsilon} + \kappa m\|\nabla v\|^2_{L^2(\omega^\varepsilon)} - \frac{\kappa^2 c}{2\alpha}(\|\nabla v\|^2_{L^2(\omega^\varepsilon)} + \|v\|^2_{\mathcal{M}^\varepsilon}).$$

Choosing $\kappa = m\alpha/c$ we have that

$$a^\varepsilon(u,v;u,v) \geq (1 - \frac{m^2\alpha}{2c})\|v\|^2_{\mathcal{M}^\varepsilon} + \frac{\alpha m^2}{2c}\big(\|u\|^2_{\mathcal{U}^\varepsilon} + \|\nabla v\|^2_{L^2(\omega^\varepsilon)}\big).$$

By choice of c, $1 - \frac{m^2\alpha}{2c} \geq \frac{1}{2}$, and the assertion holds with $C_e = \min\{1, m^2/c\}/2$. □

Lemma 3.3 (Inf-sup stability)**.** *Let $0 < \alpha \leq \alpha_0$. Then there exists a constant C_i independent of ε and α such that*

$$\sup_{(u,v)\in\mathcal{U}^\varepsilon\times\mathcal{H}^\varepsilon_\diamond} \frac{b^\varepsilon(u,v;p)}{\|(u,v)\|_\alpha} \geq C_i\|p\|_{\mathcal{H}^\varepsilon} \quad \text{for all } p \in \mathcal{H}^\varepsilon_\diamond. \tag{23}$$

Proof. Let $p \in \mathcal{H}^\varepsilon_\diamond$ be given. By Lemma A.3 the embedding $\mathcal{H}^\varepsilon_\diamond \hookrightarrow \mathcal{U}^\varepsilon$ is continuous, and thus we can choose $v = p$ and $u = -p$. Using Lemma A.5 we further obtain another constant $C > 0$, which possibly depends on α_0 but not on ε or α, such that $\|(u,v)\|_\alpha \leq C\|p\|_{\mathcal{H}^\varepsilon}$. The assertion then follows from

$$b^\varepsilon(u,v;p) \geq m\|\nabla p\|^2_{L^2(\omega^\varepsilon)} + \|p\|^2_{\mathcal{U}^\varepsilon} \geq c\|p\|^2_{\mathcal{H}^\varepsilon},$$

where we also applied (10) and Lemma A.5 with $\gamma = \gamma_H$. □

As a consequence of Brezzi's splitting theorem [5], we obtain the following result. Note that the a priori estimates derived in [5] do not use the continuity constant of b^ε.

Theorem 3.4 (Existence of saddle-points)**.** *Let $0 < \alpha \leq \alpha_0$. Then for each $f^\varepsilon \in (\mathcal{U}^\varepsilon \times \mathcal{H}^\varepsilon_\diamond)'$ and $g^\varepsilon \in (\mathcal{H}^\varepsilon_\diamond)'$ there exist a unique solution $(u^\varepsilon, v^\varepsilon) \in \mathcal{U}^\varepsilon \times \mathcal{H}^\varepsilon_\diamond$ and $p^\varepsilon \in \mathcal{H}^\varepsilon_\diamond$ of (21) and there exists a constant C_E independent of ε and α such that*

$$\alpha(\|u^\varepsilon\|^2_{\mathcal{U}^\varepsilon} + \|\nabla v^\varepsilon\|^2_{L^2(\omega^\varepsilon)}) + \|v^\varepsilon\|^2_{\mathcal{M}^\varepsilon} + \|p^\varepsilon\|^2_{\mathcal{H}^\varepsilon} \leq C_E(\|f^\varepsilon\|^2_{(\mathcal{U}^\varepsilon\times\mathcal{H}^\varepsilon_\diamond)'} + \|g^\varepsilon\|^2_{(\mathcal{H}^\varepsilon_\diamond)'}).$$

As usual $(\mathcal{U}^\varepsilon \times \mathcal{H}^\varepsilon_\diamond)'$ and $(\mathcal{H}^\varepsilon_\diamond)'$ denote the respective dual spaces of $\mathcal{U}^\varepsilon \times \mathcal{H}^\varepsilon_\diamond$ and $\mathcal{H}^\varepsilon_\diamond$, which we endow with the norms

$$\|f^\varepsilon\|_{(\mathcal{U}^\varepsilon\times\mathcal{H}^\varepsilon_\diamond)'} = \sup_{(u,v)\in\mathcal{U}^\varepsilon\times\mathcal{H}^\varepsilon_\diamond\setminus\{0\}} \frac{f^\varepsilon(u,v)}{\|(u,v)\|_\alpha}, \qquad \|g^\varepsilon\|_{(\mathcal{H}^\varepsilon_\diamond)'} = \sup_{p\in\mathcal{H}^\varepsilon_\diamond\setminus\{0\}} \frac{g^\varepsilon(p)}{\|p\|_{\mathcal{H}^\varepsilon}}.$$

3.2 Convergence and Regularization properties

In this section we will investigate the regularization properties of the diffuse domain method when used in combination with Tikhonov regularization in more detail.

Theorem 3.5 (Stability)**.** *Let $f_1, f_2 \in \mathcal{M}^\varepsilon$. Then, for C_E from Theorem 3.4, we have that*

$$\|(u_1^\varepsilon - u_2^\varepsilon, v_1^\varepsilon - v_2^\varepsilon)\|_\alpha \leq \sqrt{C_E}\|f_1 - f_2\|_{\mathcal{M}^\varepsilon},$$

where $(u_i^\varepsilon, v_i^\varepsilon)$, $i = 1, 2$, denotes the solution to (21) *with right-hand side $g^\varepsilon = 0$ and $f^\varepsilon(q, w) = \langle f_i, w\rangle_{\mathcal{M}^\varepsilon}$.*

Proof. $(u_1^\varepsilon - u_2^\varepsilon, v_1^\varepsilon - v_2^\varepsilon)$ is a solution to (21) with right-hand side $g^\varepsilon = 0$ and $f^\varepsilon(q, w) = \langle f_1 - f_2, w\rangle_{\mathcal{M}^\varepsilon}$. Since $\|f^\varepsilon\|_{(\mathcal{U}^\varepsilon \times \mathcal{H}_\diamond^\varepsilon)'} \leq \|f_1 - f_2\|_{\mathcal{M}^\varepsilon}$ the result follows directly from Theorem 3.4. □

In order show convergence of the minimizers of the diffuse Tikhonov functional as $\alpha \to 0$, we need the following technical statement, which gives some sort of compactness.

Proposition 3.6. *Let $\{(u^\varepsilon, v^\varepsilon)\} \subset \mathcal{U}^\varepsilon \times \mathcal{H}_\diamond^\varepsilon$ be a sequence such that $b^\varepsilon(u^\varepsilon, v^\varepsilon; r) = 0$ for all $r \in \mathcal{H}_\diamond^\varepsilon$ and such that there exists a constant $C > 0$ with $\|u^\varepsilon\|_{\mathcal{U}^\varepsilon} \leq C$. Then there exists a subsequence $\{v^{\varepsilon_k}\}$ of $\{v^\varepsilon\}$ and $v \in H^1(\Omega)$ such that*

$$\lim_{k\to\infty} \|\sqrt{\omega^{\varepsilon_k}}\nabla v^{\varepsilon_k} - \chi_D \nabla v\|_{L^2(\Omega)} = 0 \quad \textit{and} \quad \lim_{k\to\infty} \|v^{\varepsilon_k} - v\|_{H^1(D)} = 0.$$

Here, χ_D denotes the indicator function of D.

Proof. Using Lemma 2.4, we obtain $\|v^\varepsilon\|_{H^1(D)} \leq 2\|v^\varepsilon\|_{\mathcal{H}^\varepsilon} \leq C\|u^\varepsilon\|_{\mathcal{U}^\varepsilon} \leq C$. Thus, we can extract a subsequence $\{v^\varepsilon\}$, relabeled if necessary, such that $v^\varepsilon \rightharpoonup v$ in $H^1(D)$ as $\varepsilon \to 0$ for some $v \in H^1(D)$. Now let $\varphi \in L^2(\Omega)^n$ be arbitrary. Since $0 \leq \omega^\varepsilon \leq 1$, we obtain $|\varphi\sqrt{\omega^\varepsilon}| \leq |\varphi| \in L^2(\Omega)$. Moreover, since $\sqrt{\omega^\varepsilon} \to \chi_D$ a.e. in Ω as $\varepsilon \to 0$, we have $\varphi\sqrt{\omega^\varepsilon} \to \varphi\chi_D$ a.e. in Ω as $\varepsilon \to 0$. Hence, using dominated convergence, $\varphi\sqrt{\omega^\varepsilon} \to \varphi\chi_D$ in $L^2(\Omega)^n$, and

$$\int_D \sqrt{\omega^\varepsilon} L\nabla v^\varepsilon \cdot \varphi \,\mathrm{d}x \to \int_D L\nabla v \cdot \varphi \,\mathrm{d}x \quad \text{as } \varepsilon \to 0,$$

using the Cholesky factorization $M = L^\top L$. Since $\|\sqrt{\omega^\varepsilon} L\nabla v^\varepsilon\|_{L^2(\Omega)} \leq C\|v^\varepsilon\|_{\mathcal{H}^\varepsilon}$ is bounded (uniformly in ε), and $|(\Omega \setminus D) \cap \operatorname{supp}(\omega^\varepsilon)| \to 0$ as $\varepsilon \to 0$, absolute continuity of the integral implies

$$\int_{\Omega\setminus D} \sqrt{\omega^\varepsilon} L\nabla v^\varepsilon \cdot \varphi \,\mathrm{d}x \leq \|\sqrt{\omega^\varepsilon} L\nabla v^\varepsilon\|_{L^2(\Omega)} \|\varphi\|_{L^2((\Omega\setminus D)\cap \operatorname{supp}(\omega^\varepsilon))} \to 0,$$

as $\varepsilon \to 0$, i.e., $\sqrt{\omega^\varepsilon} L \nabla v^\varepsilon \rightharpoonup \chi_D L \nabla v$ in $L^2(\Omega)^n$ as $\varepsilon \to 0$. It remains to show that $\|\sqrt{\omega^\varepsilon} L \nabla v^\varepsilon\|_{L^2(\Omega)} \to \|\chi_D L \nabla v\|_{L^2(\Omega)}$ as $\varepsilon \to 0$. Testing $b^\varepsilon(u^\varepsilon, v^\varepsilon, r) = 0$ with $r = v^\varepsilon - v - \langle v^\varepsilon - v, 1\rangle_{\mathcal{U}^\varepsilon} / \langle 1, 1\rangle_{\mathcal{U}^\varepsilon} \in \mathcal{H}^\varepsilon_\diamond$, and applying Cauchy-Schwarz's and Young's inequality yields

$$\begin{aligned}\|\sqrt{\omega^\varepsilon} L \nabla v^\varepsilon\|^2_{L^2(\Omega)} &= \langle M \nabla v^\varepsilon, \nabla v\rangle_{\omega^\varepsilon} + \langle r, u^\varepsilon\rangle_{\mathcal{U}^\varepsilon} \\ &\leq \frac{1}{2}\|\sqrt{\omega^\varepsilon} L \nabla v^\varepsilon\|^2_{L^2(\Omega)} + \frac{1}{2}\|\sqrt{\omega^\varepsilon} L \nabla v\|^2_{L^2(\Omega)} + \|r\|_{\mathcal{U}^\varepsilon}\|u^\varepsilon\|_{\mathcal{U}^\varepsilon}.\end{aligned}$$

Since $\|r\|_{\mathcal{U}^\varepsilon} \leq 2\|v^\varepsilon - v\|_{\mathcal{U}^\varepsilon}$, this reads as

$$\|\sqrt{\omega^\varepsilon} L \nabla v^\varepsilon\|^2_{L^2(\Omega)} \leq \|\sqrt{\omega^\varepsilon} L \nabla v\|^2_{L^2(\Omega)} + 4\|v^\varepsilon - v\|_{\mathcal{U}^\varepsilon}\|u^\varepsilon\|_{\mathcal{U}^\varepsilon}. \tag{24}$$

First, we observe by using Lebesgue's dominated convergence theorem that

$$\|\sqrt{\omega^\varepsilon} L \nabla v\|^2_{L^2(\Omega)} \to \int_D M \nabla v \cdot \nabla v \,\mathrm{d}x = \|\chi_D L \nabla v\|^2_{L^2(\Omega)} \quad \text{as } \varepsilon \to 0.$$

Next, we will show that $\|v^\varepsilon - v\|_{\mathcal{U}^\varepsilon}$ vanishes as $\varepsilon \to 0$. By compactness of the embedding $H^1(D) \hookrightarrow L^2(\partial H)$, $v^\varepsilon - v \rightharpoonup 0$ in $H^1(D)$ implies $v^\varepsilon - v \to 0$ in $L^2(\partial H)$ by extracting another subsequence if necessary. Applying Theorem A.2 (i) to $v^\varepsilon - v$ we obtain

$$\|v^\varepsilon - v\|_{\mathcal{U}^\varepsilon} \leq C\sqrt{\varepsilon}\|v^\varepsilon - v\|_{\mathcal{H}^\varepsilon} + \|v^\varepsilon - v\|_{L^2(\partial H)} \to 0 \quad \text{as } \varepsilon \to 0.$$

By assumption $\{u^\varepsilon\}$ is bounded in $\mathcal{U}^\varepsilon$, and hence it follows from (24) that

$$\limsup_{\varepsilon \to 0} \|\sqrt{\omega^\varepsilon} L \nabla v^\varepsilon\|^2_{L^2(\Omega)} \leq \|\chi_D L \nabla v\|^2_{L^2(\Omega)}. \tag{25}$$

Weak lower semicontinuity of the norm further implies

$$\|\chi_D L \nabla v\|_{L^2(\Omega)} \leq \liminf_{\varepsilon \to 0} \|\sqrt{\omega^\varepsilon} L \nabla v^\varepsilon\|_{L^2(\Omega)},$$

i.e., $\|\sqrt{\omega^\varepsilon} L \nabla v^\varepsilon\|_{L^2(\Omega)} \to \|\chi_D L \nabla v\|_{L^2(\Omega)}$ as $\varepsilon \to 0$, which yields the first assertion together with the ellipticity of M (and consequent uniform bounds on the eigenvalues of L).
To show the second assertion, we infer from the first assertion that there exists another subsequence $\{\omega^\varepsilon \nabla v^\varepsilon\}$ which converges to $\chi_D \nabla v$ a.e. in Ω as $\varepsilon \to 0$. As $\omega^\varepsilon \geq 1/2$ on D we further have that for this subsequence ∇v^ε converges to ∇v a.e. in D. Moreover, with the same argument $|\nabla v^\varepsilon|^2 \leq 2\omega^\varepsilon |\nabla v^\varepsilon|^2$ on D. As $\omega^\varepsilon |\nabla v^\varepsilon|^2$ converges to $|\nabla v|^2$ in $L^1(D)$ by the first part, we obtain $\nabla v^\varepsilon \to \nabla v$ in $L^2(D)$ by dominated convergence. Compactness of the embedding $H^1(D) \hookrightarrow L^2(D)$ further yields $v^\varepsilon \to v$ in $L^2(D)$ (for a subsequence), which concludes the proof. □

The next lemma basically resembles the a priori estimates of [5]. We state it explicitly since the structure of the estimate will be of importance below. The proof is omitted.

Lemma 3.7. *Let $(u^\varepsilon_{\alpha,\delta}, v^\varepsilon_{\alpha,\delta}, p^\varepsilon_{\alpha,\delta})$ be a saddle-point of L^ε. Then there exists $C > 0$ such that*

$$\|v^\varepsilon_{\alpha,\delta} - \tilde{f}^\delta\|^2_{\mathcal{M}^\varepsilon} + \alpha\|u^\varepsilon_{\alpha,\delta}\|^2_{\mathcal{U}^\varepsilon} \leq C(\delta^2 + \alpha\|u^\dagger\|^2_{L^2(\partial H)} + \varepsilon^3\|v^\dagger\|^2_{W^{3,\infty}(D)}).$$

Using similar assumptions as in the standard inverse problem theory [12], we obtain the following convergence result.

Theorem 3.8 (Convergence)**.** *Let $\{(u^\varepsilon_{\alpha,\delta}, v^\varepsilon_{\alpha,\delta}, p^\varepsilon_{\alpha,\delta})\}$ be a sequence of saddle-points of L^ε for $\varepsilon, \alpha, \delta > 0$. If α and ε are chosen such that $\varepsilon(\alpha, \delta) \to 0$ and $\alpha(\delta) \to 0$ as $\delta \to 0$, and δ^2/α and ε^3/α are bounded. Then there exists a constant C independent of ε, δ and α such that*

$$\lim_{\delta\to 0} \|u^\varepsilon_{\alpha,\delta} - \tilde{u}^\dagger\|_{(\mathcal{H}^\varepsilon_\diamond)'} = 0, \quad \|v^\varepsilon_{\alpha,\delta} - \tilde{f}^\dagger\|_{\mathcal{M}^\varepsilon} \leq C\sqrt{\alpha} \quad \textit{and}$$
$$\|v^\varepsilon_{\alpha,\delta} - f^\dagger\|_{L^2(\partial B)} \leq C\sqrt{\alpha + \varepsilon}.$$

Proof. Applying (18) and Lemma 3.7 yields

$$\|v^\varepsilon_{\alpha,\delta} - \tilde{f}^\dagger\|_{\mathcal{M}^\varepsilon} \leq \|v^\varepsilon_{\alpha,\delta} - \tilde{f}^\delta\|_{\mathcal{M}^\varepsilon} + \|\tilde{f}^\delta - \tilde{f}^\dagger\|_{\mathcal{M}^\varepsilon} \leq C\sqrt{\alpha} \tag{26}$$

by choice of α and ε. The a priori estimate of Lemma 3.7 further asserts that

$$\|u^\varepsilon_{\alpha,\delta}\|^2_{\mathcal{U}^\varepsilon} \leq C(\frac{\delta^2}{\alpha} + \|u^\dagger\|^2_{L^2(\partial H)} + \frac{\varepsilon^3}{\alpha}\|v^\dagger\|^2_{W^{3,\infty}(D)}).$$

Since δ^2/α and ε^3/α are bounded, $\|u^\varepsilon_{\alpha,\delta}\|_{\mathcal{U}^\varepsilon}$ is bounded (uniformly in ε). By Lemma 3.6 there exists $v \in H^1(D)$ such that for a subsequence, relabeled if necessary, $v^\varepsilon_{\alpha,\delta} \to v$ in $H^1(D)$ as $\delta \to 0$. Moreover, applying (26) and Lemma A.4 (ii) yields

$$\|v^\varepsilon_{\alpha,\delta} - f^\dagger\|_{L^2(\partial B)} \leq C\|\tilde{v}^\varepsilon_{\alpha,\delta} - \tilde{f}^\dagger\|_{\mathcal{M}^\varepsilon} \leq C\sqrt{\varepsilon}\|v^\varepsilon_{\alpha,\delta}\|_{\mathcal{H}^\varepsilon} + C\sqrt{\alpha} \to 0$$

as $\delta \to 0$. In particular, $v = f^\dagger = v^\dagger \in \operatorname{ran}(F) \subset L^2(\partial B)$. Hence, there exists $u \in L^2(\partial H)$ such that $Fu = v$. Lemma 2.2 implies $u = u^\dagger$. The definition of F and unique solvability of (9) implies $v = v^\dagger$ in D. To show $u^\varepsilon_{\alpha,\delta} \to \tilde{u}^\dagger$ in $(\mathcal{H}^\varepsilon_\diamond)'$ let $w \in \mathcal{H}^\varepsilon_\diamond$, and let $v^\varepsilon \in \mathcal{H}^\varepsilon_\diamond$ denote the solution to (16) with right-hand side $\tilde{u}^\dagger$; cf. Lemma 2.5. Then

$$\begin{aligned}\langle u^\varepsilon_{\alpha,\delta} - \tilde{u}^\dagger, w\rangle_{\mathcal{U}^\varepsilon} &= \langle M\nabla(v^\varepsilon_{\alpha,\delta} - v^\varepsilon), \nabla w\rangle_{\omega^\varepsilon} \\ &= \langle M\nabla(v^\varepsilon_{\alpha,\delta} - v^\dagger), \nabla w\rangle_{\omega^\varepsilon} + \langle M\nabla(v^\dagger - v^\varepsilon), \nabla w\rangle_{\omega^\varepsilon} \\ &\leq C(\|\sqrt{\omega^\varepsilon}\nabla(v^\varepsilon_{\alpha,\delta} - v^\dagger)\|_{L^2(\Omega)} + \|v^\dagger - v^\varepsilon\|_{\mathcal{H}^\varepsilon})\|w\|_{\mathcal{H}^\varepsilon}.\end{aligned}$$

In view of Proposition 3.6 and Lemma 2.5, the right-hand side vanishes as $\delta \to 0$. The uniqueness result, Lemma 2.2, further allows to transfer the convergence to the whole sequence. □

3.3 Convergence rates

In order to show convergence rates recall that $u^\dagger$ is the minimum-norm solution of $Fu = f^\dagger$, i.e. a minimizer of

$$\min \|u\|^2_{L^2(\partial H)} \quad \text{such that } v_{|\partial B} = f^\dagger \quad \text{and} \quad b(u, v; r) = 0 \text{ for all } r \in H^1_\diamond(D).$$

The associated Lagrangian writes as

$$L(u, v, \lambda, p) = \|u\|^2_{L^2(\partial H)} - \langle v - f^\dagger, \lambda\rangle + b(u, v; p). \tag{27}$$

Assuming that there exists $(\lambda^\dagger, p^\dagger)$ such that $(u^\dagger, v^\dagger, \lambda^\dagger, p^\dagger)$ is a saddle-point of L, the following optimality conditions hold true

$$\begin{aligned}
\langle u^\dagger, h_u\rangle_{\partial H} - \langle h_u, p^\dagger\rangle_{\partial H} = 0 \quad &\text{for all } h_u \in L^2(\partial H), && (28)\\
-\langle h_v, \lambda^\dagger\rangle_{\partial B} + \langle M\nabla h_v, \nabla p^\dagger\rangle_D = 0 \quad &\text{for all } h_v \in H^1_\diamond(D), && (29)\\
\langle v^\dagger - f^\dagger, h_\lambda\rangle_{\partial B} = 0 \quad &\text{for all } h_\lambda \in L^2(\partial B), && (30)\\
b(u^\dagger, v^\dagger; h_p) = 0 \quad &\text{for all } h_p \in H^1_\diamond(D). && (31)
\end{aligned}$$

Eq. (28) implies $u^\dagger = p^\dagger$ on ∂H, where $p^\dagger$ satisfies the adjoint equation (29), i.e.

$$u^\dagger = F^* \lambda^\dagger, \tag{32}$$

which is the usual source condition. Vice versa, if (32) holds true, then (28)–(29) are satisfied, and $(u^\dagger, v^\dagger, \lambda^\dagger, p^\dagger)$ is a saddle-point of L. In order to simplify the presentation, we will assume that $n(x)$ is an eigenvector of $M(x)$ for $x \in \partial D$, i.e.

$$M(x)n(x) = a(x)n(x) \quad \text{for } x \in \partial D \tag{33}$$

for some scalar function a satisfying $m \le a(x) \le 1/m$ for all $x \in \partial D$ by (10).

Remark 3.9. *Formally, $p^\dagger$ is a solution to*

$$\begin{aligned}
-\operatorname{div}(M\nabla p^\dagger) = 0 \quad &\text{in } D,\\
n \cdot M\nabla p^\dagger = 0 \quad \text{on } \partial H, \quad &n \cdot M\nabla p^\dagger = \lambda^\dagger \quad \text{on } \partial B.
\end{aligned} \tag{34}$$

Since $n \cdot M\nabla v^\dagger = u^\dagger = p^\dagger$ on ∂H if (32) *holds, regularity assumptions on $u^\dagger$ or $v^\dagger$ can be translated to $p^\dagger$ and $\lambda^\dagger$. Similar to the assumptions on $u^\dagger$ and $v^\dagger$, we will assume that $p^\dagger \in W^{3,\infty}(D)$ in this paper. In particular, $p^\dagger$ is a strong solution to* (34).

Assuming (32) holds true, there exists a saddle-point $(u^\dagger, v^\dagger, \lambda^\dagger, p^\dagger)$ of the Lagrangian defined in (27). The error $(u^\varepsilon_{\alpha,\delta} - \tilde{u}^\dagger, v^\varepsilon_{\alpha,\delta} - v^\dagger, p^\varepsilon_{\alpha,\delta} - \alpha p^\dagger)$ satisfies the saddle-point problem (21) with right-hand side

$$f^\varepsilon(q,w) = \langle \tilde{f}^\delta, w\rangle_{\mathcal{M}^\varepsilon} - a^\varepsilon(\tilde{u}^\dagger, v^\dagger; q, w) - b^\varepsilon(q, w; \alpha p^\dagger), \tag{35}$$

$$g^\varepsilon(r) = -b^\varepsilon(\tilde{u}^\dagger, v^\dagger; r) \tag{36}$$

with $(q,w) \in \mathcal{U}^\varepsilon \times \mathcal{H}^\varepsilon_\diamond$ and $r \in \mathcal{H}^\varepsilon_\diamond$. In order to obtain error estimates we will estimate the right-hand side of the latter saddle-point problem and employ Theorem 3.4.

Lemma 3.10. *Let* (14), (33), *and* (32) *hold and let* f^ε *be defined by* (35). *Then there exists a constant* $C > 0$ *independent of* ε, α *and* δ *such that*

$$\|f^\varepsilon\|_{(\mathcal{U}^\varepsilon \times \mathcal{H}^\varepsilon_\diamond)'} \leq C(\delta + \varepsilon^{3/2}\|v^\dagger\|_{W^{3,\infty}(D)} + \varepsilon^{3/2}\alpha^{1/2}\|p^\dagger\|_{W^{3,\infty}(D)} + \alpha\|\lambda^\dagger\|_{L^2(\partial B)}).$$

Proof. Let $(q,w) \in \mathcal{U}^\varepsilon \times \mathcal{H}^\varepsilon_\diamond$. Using the source condition, i.e. $p^\dagger = u^\dagger$ on ∂H, we have that

$$f^\varepsilon(q,w) = \langle \tilde{f}^\delta - \tilde{f}^\dagger + \tilde{v}^\dagger - v^\dagger, w\rangle_{\mathcal{M}^\varepsilon} - \alpha\langle M\nabla w, \nabla p^\dagger\rangle_{\omega^\varepsilon} + \alpha\langle p^\dagger - \tilde{p}^\dagger, q\rangle_{\mathcal{U}^\varepsilon}.$$

Using (18), Cauchy-Schwarz inequality and Lemma A.4 (iii) we obtain

$$\langle \tilde{f}^\delta - \tilde{f}^\dagger, w\rangle_{\mathcal{M}^\varepsilon} + \langle v^\dagger - \tilde{v}^\dagger, w\rangle_{\mathcal{M}^\varepsilon} \leq C(\delta + \varepsilon^{3/2}\|v^\dagger\|_{W^{2,2}(D)})\|w\|_{\mathcal{M}^\varepsilon},$$

where we used $\partial_n v^\dagger = 0$ on ∂B by (33). Since $\partial_n p^\dagger = 0$ on ∂H by (33) and (34), we similarly obtain with Lemma A.4 (iii)

$$\langle p^\dagger - \tilde{p}^\dagger, q\rangle_{\mathcal{U}^\varepsilon} \leq C\varepsilon^{3/2}\|p^\dagger\|_{W^{2,2}(D)}\|q\|_{\mathcal{U}^\varepsilon}.$$

Integration by parts and $-\nabla\omega^\varepsilon = n|\nabla\omega^\varepsilon|$ yield

$$\begin{aligned}\langle M\nabla w, \nabla p^\dagger\rangle_{\omega^\varepsilon} =& -\langle \operatorname{div}(M\nabla p^\dagger), w\rangle_{\omega^\varepsilon} \\ &+ \langle n\cdot M\nabla p^\dagger, w\rangle_{\mathcal{M}^\varepsilon} + \langle n\cdot M\nabla p^\dagger, w\rangle_{\mathcal{U}^\varepsilon}\end{aligned}$$

An application of Lemma A.4 (i) yields

$$\langle \operatorname{div}(M\nabla p^\dagger), w\rangle_{\omega^\varepsilon} \leq C\varepsilon^{3/2}\|p^\dagger\|_{W^{3,\infty}(\Omega)}\|w\|_{\mathcal{H}^\varepsilon},$$

and, since $n\cdot M\nabla p^\dagger = 0$ on ∂H, Lemma A.4 (iv) gives

$$\langle n\cdot M\nabla p^\dagger, w\rangle_{\mathcal{U}^\varepsilon} \leq C\varepsilon^{3/2}\|p^\dagger\|_{W^{3,2}(\Omega;\omega^\varepsilon)}\|w\|_{\mathcal{H}^\varepsilon},$$

as well as, using $n\cdot M\nabla p^\dagger = \lambda^\dagger$ on ∂B and Lemma A.1,

$$\begin{aligned}\langle n\cdot M\nabla p^\dagger, w\rangle_{\mathcal{M}^\varepsilon} &= \langle n\cdot M\nabla p^\dagger - E_B(n\cdot M\nabla p^\dagger), w\rangle_{\mathcal{M}^\varepsilon} + \langle E_B\lambda^\dagger, w\rangle_{\mathcal{M}^\varepsilon} \\ &\leq C(\varepsilon^{3/2}\|p^\dagger\|_{W^{3,2}(\Omega;\omega^\varepsilon)}\|w\|_{\mathcal{H}^\varepsilon} + \|\lambda^\dagger\|_{L^2(\partial B)}\|w\|_{\mathcal{M}^\varepsilon}).\end{aligned}$$

Collecting the above estimates and using the definition of $\|(q,w)\|_\alpha$ yields the assertion. □

Using Lemma 3.10 and Lemma 2.5, we infer from Theorem 3.4 the following error estimate.

Theorem 3.11. *Let $0 < \alpha \leq \alpha_0$ and $\varepsilon > 0$. Moreover, let* (14), (33) *and* (32) *hold. Then there exists $C > 0$ independent of ε and α such that*

$$\begin{aligned}\alpha\|u^\varepsilon_{\alpha,\delta} - \tilde{u}^\dagger\|^2_{\mathcal{U}^\varepsilon} + \alpha\|\nabla v^\varepsilon_{\alpha,\delta} - \nabla v^\dagger\|^2_{L^2(\omega^\varepsilon)} + \|v^\varepsilon_{\alpha,\delta} - v^\dagger\|^2_{\mathcal{M}^\varepsilon} + \|p^\varepsilon_{\alpha,\delta} - \alpha p^\dagger\|^2_{\mathcal{H}^\varepsilon} \\ \leq C\big(\delta^2 + \varepsilon^3\|v^\dagger\|^2_{W^{3,\infty}(D)} + \varepsilon^3\alpha\|p^\dagger\|^2_{W^{3,\infty}(D)} + \alpha^2\|\lambda^\dagger\|^2_{L^2(\partial B)}\big).\end{aligned}$$

With an appropriate choice of ε and α in terms of δ we obtain the overall optimal order of convergence:

Corollary 3.12. *Let the assumptions of Theorem 3.11 hold true. For the a priori choice $\alpha \approx \delta$ and $\varepsilon \approx \delta^{2/3}$ we obtain the following convergence rates*

$$\begin{aligned}\|u^\varepsilon_{\alpha,\delta} - \tilde{u}^\dagger\|_{\mathcal{U}^\varepsilon} + \|\nabla v^\varepsilon_{\alpha,\delta} - \nabla v^\dagger\|_{L^2(\omega^\varepsilon)} = O(\sqrt{\delta}) \quad \text{and} \\ \|v^\varepsilon_{\alpha,\delta} - v^\dagger\|_{\mathcal{M}^\varepsilon} = O(\delta).\end{aligned} \tag{37}$$

Remark 3.13. *If $v^\dagger, p^\dagger \in W^{1,\infty}(D)$ only, we have to replace ε^3 in the previous estimates by ε, cf. Lemma A.4. The choice $\alpha \approx \delta$ and $\varepsilon \approx \delta^2$ then yields* (37).

Remark 3.14. *Assumption* (33) *can be bypassed, if one defines the extension off the interface $E_{Mn}v$ to be constant along the straight line $t \mapsto x + tM(x)n(x)$, $x \in \partial D$. Moreover, the estimates in the appendix have to be adapted in a similar way.*

We finally mention that a generalization of (32) to more general source conditions of the form $u^\dagger = (F^*F)^\mu\lambda^\dagger$ (with $0 < \mu \leq 1$) can be carried out in a similar way. The main change then concerns the last two terms on the right-hand side of the estimate in Lemma 3.10, which yield different orders in terms of α. Interestingly the optimal choice $\varepsilon^3 \approx \delta^2$ is unaffected by the specific source condition.

4 Numerical Solution

For the numerical solution we discretize the saddle-point system (21) with standard piecewise linear finite element methods on triangular grids not resolving the interface but adaptively refined based on the gradient of φ^ε. Note that this is equivalent to the optimality system for a direct finite element discretization of the minimization problem for (19). In the following we discuss some further aspects arising in the solution of the linear system.

4.1 Preconditioning of the Saddle-point System

In order to solve the saddle-point system (21) in reasonable time, we rely on efficient preconditioners. We concluded that all the constants in the stability estimates were independent of the parameter ε, cf. Lemma 3.1 and Theorem 3.4. Consequently, to obtain an ε-robust preconditioner becomes a matter of applying the proper Riesz maps, denoted by $R_{\mathcal{U}^\varepsilon} : \mathcal{U}^\varepsilon \to (\mathcal{U}^\varepsilon)'$ and $R_{\mathcal{H}^\varepsilon_\diamond} : \mathcal{H}^\varepsilon_\diamond \to (\mathcal{H}^\varepsilon_\diamond)'$. Furthermore, let us introduce the operators

$$\begin{aligned} Q^\varepsilon &: \mathcal{U}^\varepsilon \to (\mathcal{H}^\varepsilon_\diamond)', \quad u \mapsto -\langle u, w\rangle_{\mathcal{U}^\varepsilon}, \\ P^\varepsilon &: \mathcal{H}^\varepsilon_\diamond \to (\mathcal{H}^\varepsilon_\diamond)', \quad v \mapsto \langle M\nabla v, \nabla w\rangle_{\omega^\varepsilon}, \\ T^\varepsilon &: \mathcal{H}^\varepsilon_\diamond \to (\mathcal{H}^\varepsilon_\diamond)', \quad v \mapsto \langle v, w\rangle_{\mathcal{M}^\varepsilon}, \\ \tilde{T}^\varepsilon &: \mathcal{M}^\varepsilon \to (\mathcal{H}^\varepsilon_\diamond)', \quad f \mapsto \langle f, w\rangle_{\mathcal{M}^\varepsilon}, \end{aligned}$$

with $w \in \mathcal{H}^\varepsilon_\diamond$. Using these operators, we can write (21) in the form

$$\underbrace{\begin{bmatrix} \alpha R_{\mathcal{U}^\varepsilon} & 0 & [Q^\varepsilon]' \\ 0 & T^\varepsilon & [P^\varepsilon]' \\ Q^\varepsilon & P^\varepsilon & 0 \end{bmatrix}}_{\widehat{\mathcal{A}}^\varepsilon_\alpha} \underbrace{\begin{bmatrix} u^\varepsilon \\ v^\varepsilon \\ p^\varepsilon \end{bmatrix}}_{q^\varepsilon} = \underbrace{\begin{bmatrix} 0 \\ \tilde{T}^\varepsilon f \\ 0 \end{bmatrix}}_{b}, \tag{38}$$

where we have

$$\widehat{\mathcal{A}}^\varepsilon_\alpha : \mathcal{U}^\varepsilon \times \mathcal{H}^\varepsilon_\diamond \times \mathcal{H}^\varepsilon_\diamond \to (\mathcal{U}^\varepsilon)' \times (\mathcal{H}^\varepsilon_\diamond)' \times (\mathcal{H}^\varepsilon_\diamond)'. \tag{39}$$

Since this operator $\widehat{\mathcal{A}}^\varepsilon_\alpha$ maps from a (product) Hilbert space onto its dual space, Krylov subspace methods are not readily available. However, assuming that an operator $\mathcal{B}^\varepsilon : (\mathcal{U}^\varepsilon)' \times (\mathcal{H}^\varepsilon)' \times (\mathcal{H}^\varepsilon)' \to \mathcal{U}^\varepsilon \times \mathcal{H}^\varepsilon \times \mathcal{H}^\varepsilon$ is available, Krylov subspace methods can be employed to solve

$$\mathcal{B}^\varepsilon \widehat{\mathcal{A}}^\varepsilon_\alpha q^\varepsilon = \mathcal{B}^\varepsilon b.$$

To obtain an efficient solution, the preconditioner $\mathcal{B}^\varepsilon$ must be an isomorphism, see [22]. We propose to apply inverse Riesz maps to derive such a preconditioner, which lead to the preconditioned system

$$\begin{aligned} &\underbrace{\begin{bmatrix} R^{-1}_{\mathcal{U}^\varepsilon} & 0 & 0 \\ 0 & R^{-1}_{\mathcal{H}^\varepsilon_\diamond} & 0 \\ 0 & 0 & R^{-1}_{\mathcal{H}^\varepsilon_\diamond} \end{bmatrix}}_{\mathcal{B}^\varepsilon} \underbrace{\begin{bmatrix} \alpha R_{\mathcal{U}^\varepsilon} & 0 & [Q^\varepsilon]' \\ 0 & T^\varepsilon & [P^\varepsilon]' \\ Q^\varepsilon & P^\varepsilon & 0 \end{bmatrix}}_{\hat{\mathcal{A}}^\varepsilon_\alpha} \underbrace{\begin{bmatrix} u^\varepsilon \\ v^\varepsilon \\ p^\varepsilon \end{bmatrix}}_{q^\varepsilon} \\ &\qquad = \begin{bmatrix} R^{-1}_{\mathcal{U}^\varepsilon} & 0 & 0 \\ 0 & R^{-1}_{\mathcal{H}^\varepsilon_\diamond} & 0 \\ 0 & 0 & R^{-1}_{\mathcal{H}^\varepsilon_\diamond} \end{bmatrix} \underbrace{\begin{bmatrix} 0 \\ \tilde{T}^\varepsilon f \\ 0 \end{bmatrix}}_{b}. \end{aligned} \tag{40}$$

We observe that

$$\mathcal{A}_\alpha^\varepsilon = \mathcal{B}^\varepsilon \widehat{\mathcal{A}}_\alpha^\varepsilon : \mathcal{U}^\varepsilon \times \mathcal{H}_\diamond^\varepsilon \times \mathcal{H}_\diamond^\varepsilon \to \mathcal{U}^\varepsilon \times \mathcal{H}_\diamond^\varepsilon \times \mathcal{H}_\diamond^\varepsilon, \tag{41}$$

and consequently, since $\mathcal{A}_\alpha^\varepsilon$ is a symmetric indefinite operator, the MINRES algorithm can be applied to solve the optimality system.

Remark 4.1. *For our numerical examples we will use a norm induced by the inner product*

$$\langle M\nabla v, \nabla v\rangle_{\omega^\varepsilon} + \langle v, v\rangle_{\omega^\varepsilon} \tag{42}$$

on $\mathcal{H}_\diamond^\varepsilon$. This influences the preconditioner $\mathcal{B}^\varepsilon$, resulting in a slightly different stiffness matrix from the discretization of the Riesz map $R_{\mathcal{H}_\diamond^\varepsilon}$. From a numerical investigation, this gave better iteration counts, and we therefore apply this alternative norm in the numerical section.

4.2 Spectrum of the preconditioned system

Operators similar to $\mathcal{A}_\alpha^\varepsilon$ were thoroughly analyzed in [24]. Under given assumptions, an efficient and robust solution of the saddle-point system (40) can be guaranteed. More specifically, the authors of [24] show that for a sound discretization of $\mathcal{A}_\alpha^\varepsilon$ defined in (40)-(41), the spectrum of the associated discretized operator $\mathcal{A}_\alpha^{\varepsilon,h}$ satisfied

$$\mathrm{sp}(\mathcal{A}_\alpha^{\varepsilon,h}) \subset [-b,-a] \cup [c\alpha, 2\alpha] \cup \{\tau_1, \tau_2, ..., \tau_{N(\alpha)}\} \cup [a,b], \tag{43}$$

where $N(\alpha) = O(\ln(\alpha^{-1}))$ and the constants a, b, c are independent of α (and here also of ε).
To guarantee this spectrum, the following assumptions must be satisfied:

$\mathcal{A}\mathbf{1}$: $P^\varepsilon : \mathcal{H}_\diamond^\varepsilon \to (\mathcal{H}_\diamond^\varepsilon)'$ is bounded, linear, and invertible.

$\mathcal{A}\mathbf{2}$: $Q^\varepsilon : \mathcal{U}_\beta^\varepsilon \to (\mathcal{H}_\diamond^\varepsilon)'$ is bounded and linear.

$\mathcal{A}\mathbf{3}$: $T^\varepsilon : \mathcal{H}_\diamond^\varepsilon \to (\mathcal{H}_\diamond^\varepsilon)'$ is bounded and linear.

$\mathcal{A}\mathbf{4}$: The operator equation (17) is ill-posed.

Assumptions $\mathcal{A}\mathbf{1}$-$\mathcal{A}\mathbf{4}$ follow immediately from the analysis in Section 3.

4.3 Implementation

We implemented the code using cbc.block, which is a FEniCS-based Python implemented library for block operators. See [21] for details. The PyTrilinos package was used to compute an approximation of the preconditioner $\mathcal{B}^\varepsilon$ in (40). We approximated $\mathcal{B}^\varepsilon$ using AMG with a symmetric Gauß-Seidel smoother with three smoothing sweeps. All tables containing iteration counts for the MINRES method were generated with this approximate

preconditioner. On the other hand, the eigenvalues of $\mathcal{A}_\alpha^\varepsilon = \mathcal{B}^\varepsilon \widehat{\mathcal{A}}_\alpha^\varepsilon$ were computed with the *exact* preconditioner $\mathcal{B}^\varepsilon$ in Octave. The MINRES iteration process was stopped as soon as

$$\frac{\|r_n\|}{\|r_0\|} = \frac{\|\mathcal{B}^\varepsilon[\hat{\mathcal{A}}_\alpha^\varepsilon q_n - b]\|_{\mathcal{U}^\varepsilon \times \mathcal{H}^\varepsilon \times \mathcal{H}^\varepsilon}}{\|\mathcal{B}^\varepsilon[\hat{\mathcal{A}}_\alpha^\varepsilon q_0 - b]\|_{\mathcal{U}^\varepsilon \times \mathcal{H}^\varepsilon \times \mathcal{H}^\varepsilon}} < \rho. \tag{44}$$

Here, ρ is a small positive parameter. The exact data $u^\dagger$ was computed from an appropriate source condition, i.e. $F^* w = u^\dagger$, for some $w \in L^2(\partial B)$. Then, we computed $Fu^\dagger = f^\dagger$. Noise was then added to $f^\dagger$, and the noisy data was extended to $\mathrm{supp}(\gamma_B |\nabla \omega^\varepsilon|)$ by the extension operator E_B, see Section 2.3.2.

4.4 Examples

In our simulations, we use a "circle in circle" domain. The domain D is defined as

$$D = \{(x, y) \in \mathbb{R}^2 : 0.3 < \sqrt{x^2 + y^2} < 1\}.$$

The diffuse domain D_ε is then simply the scaling

$$D_\varepsilon = \{(x, y) \in \mathbb{R}^2 : 0.3 - \varepsilon < \sqrt{x^2 + y^2} < 1 + \varepsilon\}.$$

Furthermore, the conductivity tensor $M(x, y)$ is defined as

$$M = \bar{L} \Sigma \bar{L}^\top,$$

where

$$\bar{L} = \frac{1}{\|(x, y)\|} \begin{bmatrix} y & x \\ -x & y \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 0.3 \end{bmatrix}.$$

One easily verifies that (33) holds for this choice of M. In Table 1, we see the iteration numbers for different values of α and ε. As expected, there is no dependency on the diffuse domain parameter ε, cf. Section 4.2. Furthermore, for the regularization parameter α, we get the expected logarithmic growth in iteration numbers when $\alpha \to 0$. For example, when $\varepsilon = 2^{-6}$, the growth is well modeled by the function

$$\alpha \mapsto 55 - 24 \log_{10}(\alpha).$$

Figure 1 shows the eigenvalues of $\mathcal{A}_\alpha$. The band structure is in accordance with the analysis in [24], with three bands of eigenvalues, and a limited number of isolated eigenvalues.
We recall Assumption $\mathcal{A}\mathbf{4}$, i.e. that the operator equation (17) is ill-posed. In Figure 2, logarithmic plots of the absolute values of the eigenvalues of $\mathcal{A}_0^\epsilon$ are displayed. The clustering of eigenvalues around 0 is an effect of the ill-posed nature of (17).

$\varepsilon \setminus \alpha$	1	.1	.01	.001	.0001
2^{-2}	57	100	143	186	238
2^{-3}	57	91	126	157	195
2^{-4}	64	102	126	144	183
2^{-5}	57	83	115	143	159
2^{-6}	55	79	105	123	155

Table 1: The number of MINRES iteration required to solve the discretized system associated with (40). The stopping criterion $\rho = 10^{-10}$, see (44).

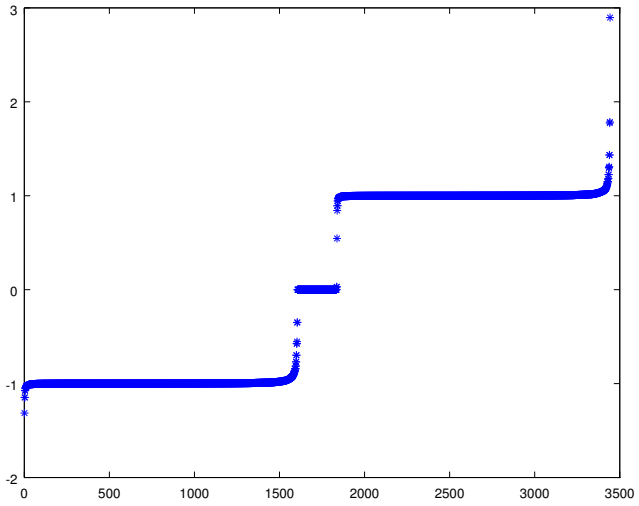
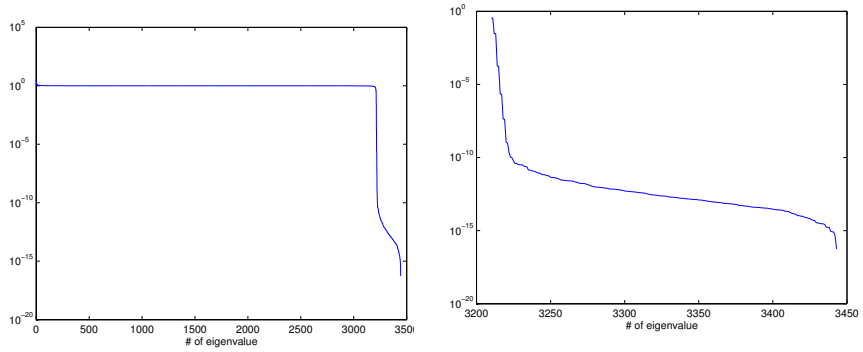


Figure 1: Plot of the eigenvalues associated with $\mathcal{A}_\alpha^\varepsilon$ in Example 1. Here $\alpha = 10^{-4}$ and $\varepsilon = 0.125$. The eigenvalues are computed on a course mesh with 1 605 vertices.

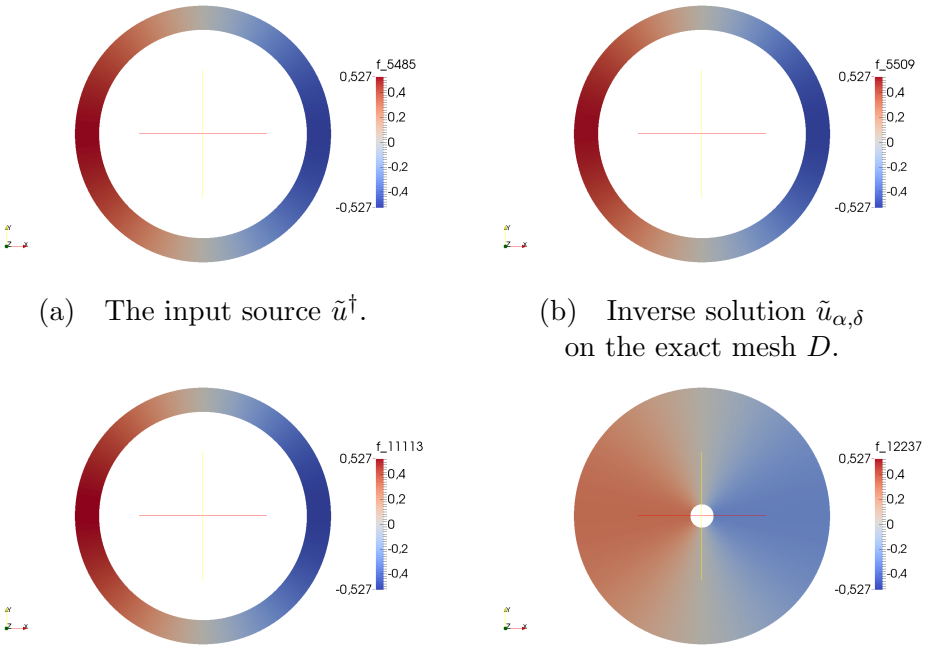
From a practical point of view, we are concerned with the performance of the diffuse domain method in comparison to the standard inverse formulation, i.e. with the optimization performed on the exact domain. We will compare the solutions both visually and in norm sense.
In Figure 3, the exact source function is displayed along with inverse solutions on both the exact and diffuse mesh. Similar comparisons are displayed in Figures 4 and 5 for the state and adjoint functions, respectively. The functions defined on a surface, i.e. either on ∂H or ∂B, are extended by the appropriate constant extension operator, see Section 2.3.2.
For the control functions, the inverse solution $u_{\alpha,\delta}$ displayed in Figure 3b) is visually identical to the exact source function $u^\dagger$. These are also visually identical to $u_{\alpha,\delta}^\varepsilon$ displayed in Figure 3c), where $\varepsilon = 0.03125 = \sqrt{\delta}$. With a larger choice of ε, however, the solution is quite different from the source $u^\dagger$, see Figure 3d) where $\varepsilon = 1/4$. If we consider the state functions, the choice of ε is less important. All solutions displayed in Figure 4 are basically



(a) All eigenvalues.

(b) Zoomed in on the smallest eigenvalues.

Figure 2: Logarithmic plots of the absolute values of the eigenvalues of $\mathcal{A}_0^\varepsilon$.



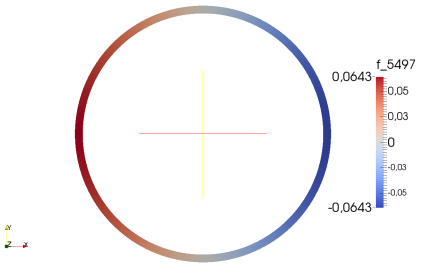
(a) The input source $\tilde{u}^\dagger$.

(b) Inverse solution $\tilde{u}_{\alpha,\delta}$ on the exact mesh D.

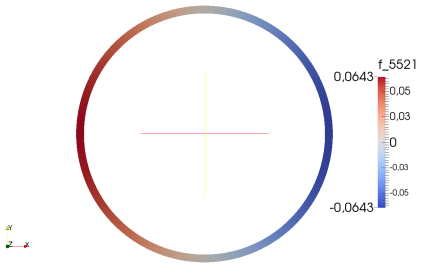
(c) Diffuse solution $u^\varepsilon_{\alpha,\delta}$ for $\varepsilon = 0.03125 = \delta^{1/2}$.

(d) Diffuse solution $u^\varepsilon_{\alpha,\delta}$ for $\varepsilon = 1/4$.

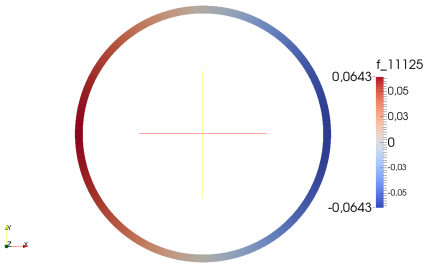
Figure 3: A comparison of different control functions and the input source in a). In a) and b), the control is only defined on ∂H, so we therefore applied the constant extension E_H for the visualization, see Section 2.3.2. In b), c) and d), $\delta = 2^{-10}$ and $\alpha = \delta/2$.



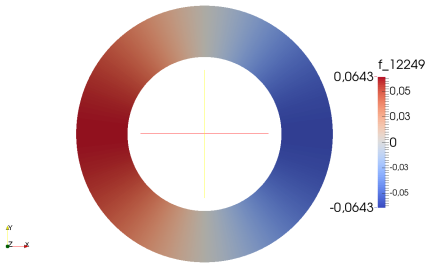
(a) The exact data $\tilde{f}^\dagger$.



(b) The state $\tilde{f}_{\alpha,\delta} = E_B(v_{\alpha,\delta|\partial B})$ on the exact mesh D.



(c) Diffuse state $v^\varepsilon_{\alpha,\delta}$ for $\varepsilon = 0.03125 = \delta^{1/2}$.



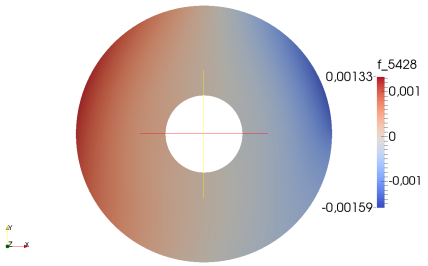
(d) Diffuse state $f^\varepsilon_{\alpha,\delta}$ for $\varepsilon = .25$.

Figure 4: A comparison of different state functions and the exact data in a). In a) and b), the state is only defined on ∂B, so we therefore applied the constant extension E_B for the visualization, see Section 2.3.2. In b), c) and d), $\delta = 2^{-10}$ and $\alpha = \delta/2$

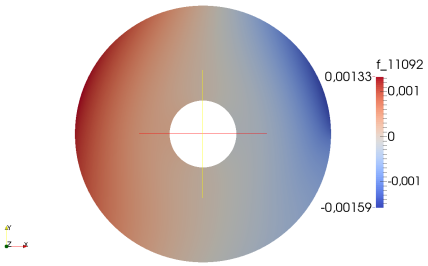
identical from a visual perspective. For the adjoint functions, there seem to be some visual difference between $p_{\alpha,\delta}$ and $p^\varepsilon_{\alpha,\delta}$, i.e. for the adjoint on the exact mesh and on the diffuse mesh for $\varepsilon = 0.03125$, but the order of magnitude of these functions is only 10^{-3}.
The final issue we will investigate numerically is the convergence rates of

$$\|u^\varepsilon_{\alpha,\delta} - \tilde{u}^\dagger\|_{\mathcal{U}^\varepsilon},$$

for choices of $\alpha = C\delta^\mu$ and $\varepsilon = c\delta^\nu$. In Figure 6 we see convergence rates for the choice $\alpha = \delta/2$. In a), the convergence rate on the exact mesh is displayed. The rate seems, on average, to be of order $O(\delta^{1/2})$, but it is quite inconsistent from step to step. This leads us to believe that a stronger source condition holds true and better convergence rates may be obtained, see Section 3.3. If the smoother source condition is satisfied, we can choose $\alpha = C\delta^{2/3}$. The convergence rates for this choice of α is displayed in Figure 7. In a), we now see a much more consistent convergence rate of order $O(\delta^{2/3})$.



(a) The adjoint $p_{\alpha,\delta}$ on the exact mesh D.

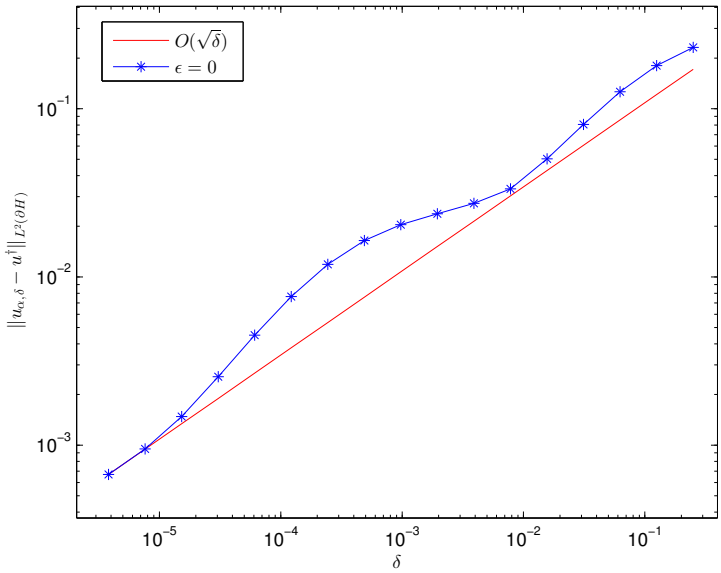


(b) The diffuse adjoint $p^{\delta}_{\alpha,\varepsilon}$ for $\varepsilon = .03125 = \delta^{1/2}$.

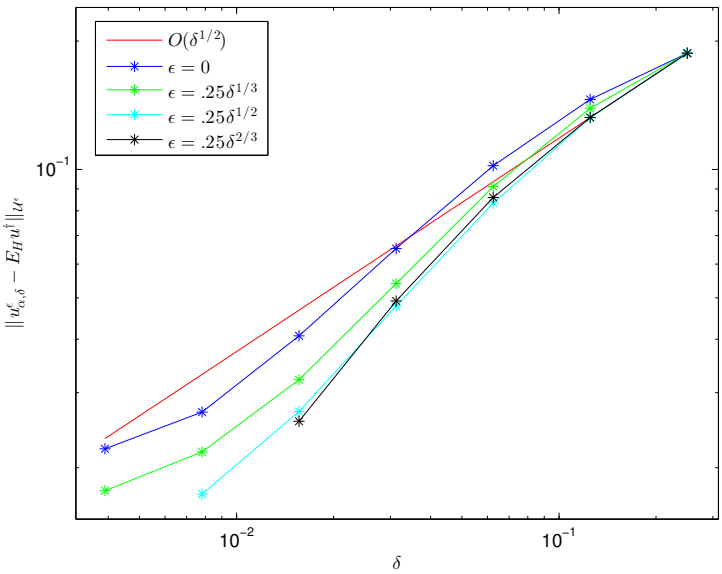
Figure 5: A comparison of the adjoint on the exact mesh and the diffuse mesh. Here, $\delta = 2^{-10}$ and $\alpha = \delta/2$.

For the convergence rates associated with the diffuse domain method, we have more inconsistent rates. Generally, the convergence rates can only be guaranteed for small choices of δ and ε, and particularly the latter is difficult to handle numerically, due to mesh limitations on standard computers. However, we see in Figure 6b) that the choices $\varepsilon = \delta^{1/2}/4$ and $\varepsilon = \delta^{2/3}/4$ yield roughly the same convergence rate, while $\varepsilon = \delta^{1/3}/4$ yields a worse rate.

For the case $\alpha = C\delta^{2/3}$, displayed in Figure 7, the numerics become more challenging. We observe from the rates associated with the inverse solutions on the exact mesh that we only obtain the theoretical convergence $\|u_{\alpha,\delta} - u^{\dagger}\|_{L^2(\partial H)} = O(\delta^{2/3})$ for small values of δ. Hence, choosing $\varepsilon = \delta^{\nu}$ might be numerically challenging for these values of δ. However, the constant in Theorem 3.11 is not explicit, and we therefore select heuristically C in $\varepsilon = C\delta^{\nu}$. From Figure 7b), we observe that the choice $\varepsilon = 35\delta^{2/3}$ yields a better rate than choosing $\varepsilon = 10\delta^{1/2}$, which again yields a better rate than $\varepsilon = 2.8\delta^{1/3}$. Furthermore, for the smallest noise values, the convergence rate associated with the choices $\varepsilon = 35\delta^{2/3}$ and $\varepsilon = 10\delta^{1/2}$ actually seems to be of order $O(\delta^{2/3})$, which is the optimal rate from standard theory, see [12]. The choice $\varepsilon = C\delta^{1/2}$ is better than our theory suggests. Roughly, this may be explained as follows. Measuring in a norm similar to a weighted $W^{1,1}$-norm gives approximations of order ε^2 instead of $\varepsilon^{3/2}$, see [7] and Theorem A.2. Using this in Theorem 3.11, the optimal choice in Corollary 3.12 is actually $\varepsilon \approx \delta^{1/2}$. As for coarse discretizations all norms are equivalent with moderate constants this may explain the observed behavior.

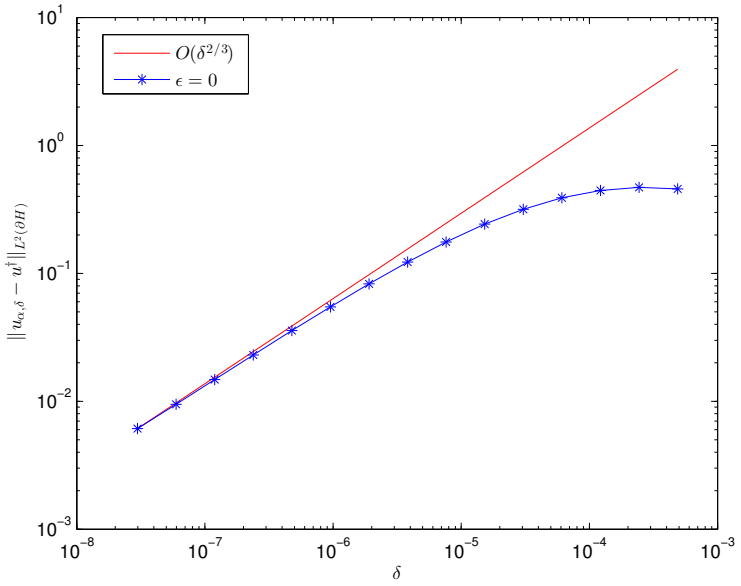


(a) Convergence rate for the control on the exact mesh D.

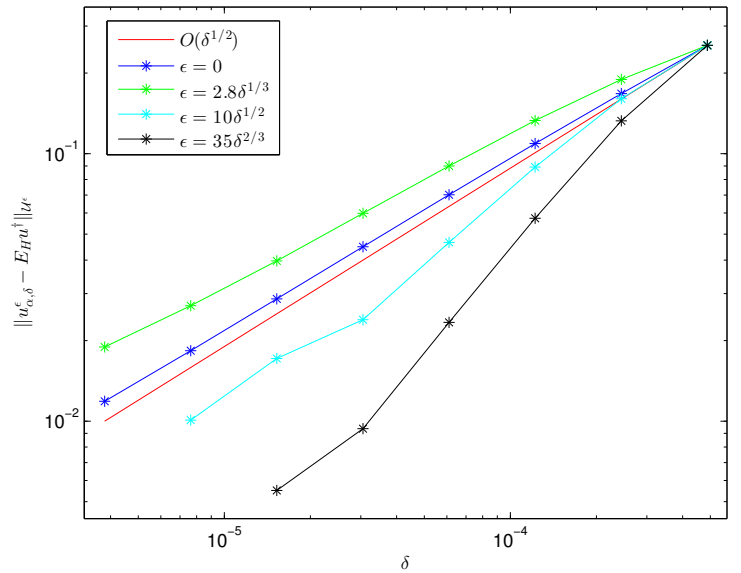


(b) Convergence rate for the control on the diffuse mesh D_ε, with $\varepsilon = .25\delta^\nu$ where $\nu = \{1/3, 1/2, 2/3\}$.

Figure 6: A log-log plot of the convergence rates for different choices of diffuse domain parameter ε. In both subplots we see the actual convergence rates (experimental), compared to the theoretical rate of order $O(\varepsilon^{1/2})$. Here, $\alpha = \delta/2$.



(a) Convergence rate for the control on the exact mesh D for $\alpha = 2\delta^{2/3}$.



(b) Convergence rate for the control on the diffuse mesh D_ε, with $\varepsilon = C\delta^\nu$ and for $\alpha = \delta^{2/3}$.

Figure 7: A log-log plot of the convergence rates for different choices of diffuse domain parameter ε. In both subplots we see the actual convergence rates (experimental), compared to the theoretical rate of order $O(\delta^{2/3})$. All errors in b) are scaled to be equal at the largest noise value. The notation $\varepsilon = 0$ means computations on the exact mesh, i.e. as in a).

5 Discussion and conclusions

We applied a diffuse domain method to variational regularization methods. This allowed us to handle complex geometries in a computationally efficient way. The additional error introduced by the diffuse domain method can be made arbitrarily small such that the overall error in the method is dominated by modeling errors and measurement noise. As a model problem we chose ECG inversion for which we could show that Tikhonov regularization is indeed a regularization method. Extensions to other inverse problems governed by an elliptic partial differential equation of second order seem to be straightforward. The main difference to standard Tikhonov regularization in Hilbert spaces, where simple operator perturbations can be handled in a straightforward manner, is the choice of topology which depends on ε, the parameter in the diffuse domain method. As this topology is weaker than the standard Hilbert space norm, we could show convergence in a dual norm only. A key ingredient for our convergence result is the reformulation of Tikhonov regularization as a constraint optimization problem, which gives additional control over the state, which in turn gave some compactness. Under the usual source conditions we could prove convergence rates in the stronger standard Hilbert space norm when an a priori parameter choice rule is used. Using the methods present here, it should be possible to analyze also other parameter choice rules, and the use of nonlinear forward problems should also be feasible. Extending the results of [7] to parabolic problems, one can also deal with time-dependent inverse problems. Here, the diffuse domain method is particularly suited when dealing with time-dependent geometries as e.g. a beating heart. Another interesting point, which is not in the scope of this paper, is how errors in the distance function will influence the diffuse domain method. On the continuous level noisy distance functions will lead to rough surfaces and new challenges come up. Of particular interest is the case when only finitely many measurements, and hence measurement locations, are available, which makes it necessary to construct a distance function in a way that the noise is not dominant.

Acknowledgements

MB and MS acknowledge support by ERC via Grant EU FP 7 - ERC Consolidator Grant 615216 LifeInverse. MB acknowledges support by the German Science Foundation DFG via EXC 1003 Cells in Motion Cluster of Excellence, Münster, Germany. OLE acknowledges support by DAAD for his one year research stay at WWU Münster.

References

[1] R. A. Adams. *Sobolev spaces*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975. Pure and Applied Mathematics, Vol. 65.

[2] N. D. Alikakos, P. W. Bates, and X. Chen. Convergence of the Cahn-Hilliard equation to the Hele-Shaw model. *Archive for rational mechanics and analysis*, 128(2):165–205, 1994.

[3] P. Bastian and C. Engwer. An unfitted finite element method using discontinuous Galerkin. *Internat. J. Numer. Methods Engrg.*, 79(12):1557–1576, 2009.

[4] M. Bertalmio, F. Mémoli, L.-T. Cheng, G. Sapiro, and S. Osher. Variational problems and partial differential equations on implicit surfaces: Bye bye triangulated surfaces? In *Geometric Level Set Methods in Imaging, Vision, and Graphics*, pages 381–397. Springer, 2003.

[5] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge*, 8(R-2):129–151, 1974.

[6] M. Burger. Finite element approximation of elliptic partial differential equations on implicit surfaces. *Computing and visualization in science*, 12(3):87–100, 2009.

[7] M. Burger, O. L. Elvetun, and M. Schlottbom. Analysis of the diffuse domain method for second order elliptic boundary value problems. *submitted*, 2014.

[8] G. Caginalp. Stefan and Hele-Shaw type models as asymptotic limits of the phase-field equations. *Physical Review A*, 39(11):5887, 1989.

[9] J. W. Cahn and J. E. Hilliard. Free energy of a nonuniform system. i. interfacial free energy. *The Journal of chemical physics*, 28(2):258–267, 1958.

[10] K. Deckelnick, C. M. Elliott, and V. Styles. Double obstacle phase field approach to an inverse problem for a discontinuous diffusion coefficient. Technical Report 1504.01935, arxiv, 2015.

[11] M. C. Delfour and J.-P. Zolésio. *Shapes and geometries*, volume 22 of *Advances in Design and Control*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 2011. Metrics, analysis, differential calculus, and optimization.

[12] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[13] S. Ghosh and Y. Rudy. Application of l1-norm regularization to epicardial potential solution of the inverse electrocardiography problem. *Annals of biomedical engineering*, 37(5):902–912, 2009.

[14] R. Glowinski, T.-W. Pan, and J. Périaux. A fictitious domain method for Dirichlet problem and applications. *Comput. Methods Appl. Mech. Engrg.*, 111(3-4):283–303, 1994.

[15] W. Hackbusch and S. A. Sauter. Composite finite elements for the approximation of PDEs on domains with complicated micro-structures. *Numer. Math.*, 75(4):447–472, 1997.

[16] D. Khoury. Use of current density an the regularization of the inverse problem of electrocardiography. In *Engineering in Medicine and Biology Society, 1994. Engineering Advances: New Opportunities for Biomedical Engineers. Proceedings of the 16th Annual International Conference of the IEEE*, pages 133–134. IEEE, 1994.

[17] K. Y. Lervag and J. Lowengrub. Analysis of the diffuse-domain method for solving PDEs in complex geometries. *arxiv:1407.7480v1*, 2014.

[18] R. J. LeVeque and Z. L. Li. The immersed interface method for elliptic equations with discontinuous coefficients and singular sources. *SIAM J. Numer. Anal.*, 31(4):1019–1044, 1994.

[19] X. Li, J. Lowengrub, A. Rätz, and A. Voigt. Solving PDEs in complex geometries: a diffuse domain approach. *Commun. Math. Sci.*, 7(1):81–107, 2009.

[20] F. Liehr, T. Preusser, M. Rumpf, S. Sauter, and L. O. Schwen. Composite finite elements for 3D image based computing. *Comput. Vis. Sci.*, 12(4):171–188, 2009.

[21] K. A. Mardal and J. B. Haga. Block preconditioning of systems of PDEs. In A. Logg, K. A. Mardal, and G. Wells, editors, *Automated Solution of Differential Equations by the Finite Element Method*, pages 643–654. Springer, 2012.

[22] K. A. Mardal and R. Winther. Preconditioning discretizations of systems of partial differential equations. *Numerical Linear Algebra with Applications*, 18(1):1–40, 2011.

[23] C. Miranda. *Partial differential equations of elliptic type.* Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 2. Springer-Verlag, New

York-Berlin, 1970. Second revised edition. Translated from the Italian by Zane C. Motteler.

[24] B. F. Nielsen and K. A. Mardal. Analysis of the Minimal Residual Method applied to ill-posed optimality systems. *SIAM Journal on Scientific Computing*, 35(2):A785–A814, 2013.

[25] A. Rätz, A. Voigt, et al. Pde's on surfaces—a diffuse interface approach. *Communications in Mathematical Sciences*, 4(3):575–590, 2006.

[26] R. Throne and L. Olson. A comparison of spatial regularization with zero and first order Tikhonov regularization for the inverse problem of electrocardiography. In *Computers in Cardiology 2000*, pages 493–496. IEEE, 2000.

A Basic Properties of Diffuse Approximations

In this appendix we collect and extend some results of [7]. We let E be one of the extensions E_B or E_H defined in Section 2.3.2 and γ be one of the weighting functions γ_B or γ_H, and assume that ε_0 is sufficiently small. Moreover let $\Gamma = \partial D \cap \operatorname{supp}(\gamma)$. The constants C are independent of ε. For $t \in (-\varepsilon, \varepsilon)$, we define the mapping $\Phi_t(x) = x + tn(x)$, $x \in \partial D$, and note that $\Phi_t(\partial D) = \{x \in \Omega : d_D(x) = t\}$. Moreover, cf. [7, Eq. (9)],

$$\lim_{t\to 0} \sup_{x\in\Gamma} |\det D\Phi_t(x) - 1 - t\Delta d_D(x)| = 0. \tag{45}$$

For any integrable v the transformation formula implies

$$\int_\Omega v(x)|\nabla\omega^\varepsilon|\gamma \,\mathrm{d}x = \frac{1}{2\varepsilon}\int_\Gamma \int_{-\varepsilon}^{\varepsilon} v(x+tn(x))|\det D\Phi_t(x)| \,\mathrm{d}t \,\mathrm{d}\sigma(x). \tag{46}$$

Let us begin with deriving some basic properties of the extensions constant off the interface defined in Section 2.3.2.

Lemma A.1. *There exists constant $c(\varepsilon), C(\varepsilon) > 0$ such that for any $v \in L^2(\Gamma)$*

$$c(\varepsilon)\|v\|_{L^2(\Gamma)} \le \|Ev\|_{L^2(\gamma|\nabla\omega^\varepsilon|)} \le C(\varepsilon)\|v\|_{L^2(\Gamma)}$$

and $c(\varepsilon) \to 1$ and $C(\varepsilon) \to 1$ as $\varepsilon \to 0$.

Proof. According to (46) and $(Ev)(x + tn(x)) = v(x)$, $x \in \Gamma$, we have

$$\int_\Omega |E_B f(x)|^2 |\nabla\omega^\varepsilon|\gamma_B \,\mathrm{d}x = \frac{1}{2\varepsilon}\int_{\partial B} |f(x)|^2 \int_{-\varepsilon}^{\varepsilon} \det D\Phi_t(x) \,\mathrm{d}t \,\mathrm{d}\sigma(x),$$

and the assertion follows from (45). □

Lemma A.1 implies that E_B and E_H are bounded, injective and have closed range.
The next issue, concerns the approximation of diffuse integrals. We set

$$\Gamma_t = \{x \in \Omega : \operatorname{dist}(x, \Gamma) < t\}.$$

The following is a central estimate.

Theorem A.2. *Let $1 \leq p < \infty$. There exists a constant $C > 0$ such that (i) if $v \in W^{1,p}(\Omega; \omega^\varepsilon)$, then*

$$\|v\|^p_{L^p(\Gamma_\varepsilon; |\nabla \omega^\varepsilon| \gamma)} \leq C(\|v\|^p_{L^p(\Gamma)} + \varepsilon^{p-1} \|\partial_n v\|^p_{L^p(\Gamma_\varepsilon; \omega^\varepsilon)}).$$

(ii) if $v \in W^{2,p}(\Omega; \omega^\varepsilon)$, then

$$\|v\|^p_{L^p(\Gamma_\varepsilon; |\nabla \omega^\varepsilon| \gamma)} \leq C(\|v\|^p_{L^p(\Gamma)} + \varepsilon^{p} \|\partial_n v\|^p_{L^p(\Gamma)} + \varepsilon^{2p-1} \|\partial_n^2 v\|^p_{L^p(\Gamma_\varepsilon; \omega^\varepsilon)}).$$

Proof. (i) Using the basic inequality $(a+b)^p \leq 2^{p-1}(|a|^p + |b|^p)$, $a, b \in \mathbb{R}$, we obtain by using the fundamental theorem of calculus and Hölders inequality

$$|v(x + tn(x))|^p \leq 2^{p-1}(|v(x)|^p + |t|^{p-1} \int_{-|t|}^{|t|} |\partial_n v(x + sn(x))|^p \, \mathrm{d}s).$$

Using the latter in (46) and using (45), we obtain

$$\|v\|^p_{L^p(\Omega; |\nabla \omega^\varepsilon| \gamma)|)} \leq 2^{p-1}(\|v\|^p_{L^p(\Gamma)} + \varepsilon^{p-2} \int_\Gamma \int_0^\varepsilon \int_{-t}^t |\partial_n v(\Phi_s(x))|^p \, \mathrm{d}s \, \mathrm{d}t \, \mathrm{d}\sigma).$$

Using Fubini's theorem we further may write

$$\frac{1}{\varepsilon} \int_\Gamma \int_0^\varepsilon \int_{-t}^t |\partial_n v(\Phi_s(x))|^p \, \mathrm{d}s \, \mathrm{d}t \, \mathrm{d}\sigma \leq C \frac{1}{\varepsilon} \int_0^\varepsilon \int_{\Gamma_t} |\partial_n v(x)|^p \, \mathrm{d}x \, \mathrm{d}t.$$

As in [7, Section 5.1] using the transformation $s = -S(t/\varepsilon)$, one completes the proof showing

$$\frac{1}{\varepsilon} \int_0^\varepsilon \int_{\Gamma_t} |\partial_n v(x)|^p \, \mathrm{d}x \, \mathrm{d}t \leq \int_{\Gamma_\varepsilon} |\partial_n v(x)|^p \omega^\varepsilon \, \mathrm{d}x.$$

(ii) Applying twice the fundamental theorem of calculus yields

$$v(x + tn(x)) = v(x) + t\partial_n v(x) + \int_0^t \int_0^s \partial_n^2 v(x + rn(x)) \, \mathrm{d}r \, \mathrm{d}s.$$

The proof is then finished with similar arguments as in (i). □

With the usual modifications one shows that Theorem A.2 also holds for $p = \infty$. We start with a diffuse trace lemma, cf. [7, Theorem 4.2]. We give a different proof.

Lemma A.3. *There exists a constant $C > 0$ such that*

$$\|v\|_{L^2(\gamma|\nabla\omega^\varepsilon|)} \leq C\|v\|_{\mathcal{H}^\varepsilon} \text{ for all } v \in \mathcal{H}^\varepsilon. \tag{47}$$

Proof. The usual trace theorem [1] assures that $\|v\|_{L^2(\Gamma)} \leq C\|v\|_{H^1(D)} \leq C\|v\|_{\mathcal{H}^\varepsilon}$. The result then follows from Theorem A.2 (i). □

Operator perturbations induced by the diffuse integrals can be treated using the following.

Lemma A.4. *Let $1 \leq p \leq \infty$ and $v \in W^{k,p}(\Omega, \omega^\varepsilon)$, $k \in \{0, 1, 2\}$. Then there exists a constant $C > 0$ independent of ε such that*
(i) if $k \leq 1$ there holds

$$|\int_\Omega v\omega^\varepsilon \,\mathrm{d}x - \int_D v \,\mathrm{d}x| \leq C\varepsilon^{1+k-\frac{1}{p}}\|v\|_{W^{k,p}(\Omega;\omega^\varepsilon)},$$

(ii) if $k = 1$, then

$$\|v - Ev\|_{L^p(|\nabla\omega^\varepsilon|\gamma)} \leq C\varepsilon^{1-\frac{1}{p}}\|v\|_{W^{1,p}(\Omega;\omega^\varepsilon)},$$

(iii) if $k = 2$, then

$$\|v - Ev\|_{L^p(\gamma|\nabla\omega^\varepsilon|)} \leq C(\varepsilon\|\partial_n v\|_{L^p(\Gamma)} + \varepsilon^{2-\frac{1}{p}}\|\partial_n^2 v\|_{L^p(\Gamma_\varepsilon;\omega^\varepsilon)}).$$

(iv) if $k = 2$, $v = 0$ on Γ and $w \in W^{1,2}(\Omega;\omega^\varepsilon)$, then

$$|\int_\Omega vw|\nabla\omega^\varepsilon|\gamma \,\mathrm{d}x| \leq C\varepsilon^{\frac{3}{2}}\|v\|_{W^{2,2}(\Omega;\omega^\varepsilon)}\|w\|_{W^{1,2}(\Omega;\omega^\varepsilon)}.$$

Proof. Assertions (i) and (iv) are proven in [7, Theorem 5.1, Theorem 5.2, Theorem 5.6]. To prove (ii) we apply Theorem A.2 (i) to $v - Ev$. As $v - Ev = 0$ on Γ and $\partial_n(v - Ev) = \partial_n v$, we obtain

$$\|v - Ev\|_{L^p(\Gamma_\varepsilon;|\nabla\omega^\varepsilon|\gamma)} \leq C\varepsilon^{1-\frac{1}{p}}\|\partial_n v\|_{L^p(\Gamma_\varepsilon;\omega^\varepsilon)}.$$

This yields the assertion. (iii) is a direct consequence of Theorem A.2 (ii). □

A further tool in studying the diffuse domain method is the following lemma [7, Lemma 4.9].

Lemma A.5 (Poincaré-Friedrichs-type inequality)**.** *There exists a constant $C > 0$ such that*

$$\|v\|^2_{\mathcal{H}^\varepsilon} \leq C(\|\nabla v\|^2_{L^2(\Omega;\omega^\varepsilon)} + \|v\|^2_{L^2(\Omega;\gamma|\nabla\omega^\varepsilon|)}) \quad \text{for all } v \in \mathcal{H}^\varepsilon. \tag{48}$$